ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

**Τίτλος διπλωματική εργασίας**

***«Επιδημιολογική και Κλινική Μελέτη γυναικών με κύστη ωοθήκης σε Νοσηλευτικό Ίδρυμα της Ελλάδας: Συστηματική ανασκόπηση και αναδρομική μελέτη με χρήση R studio.»***

**Ονοματεπώνυμο**
**Σταθοπούλου Καλλιόπη-Μαρία**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**
**Επιβλέπων**
**Πλαγιανάκος Βασίλειος**

Λαμία, 2019

UNIVERCITY OF THESSALY

SCHOOL OF SCIENCE

INFORMATICS AND COMPUTATIONAL BIOMEDICINE

**Title**

# «Epidemiological and clinical study of women with ovarian cyst treted in a tertiary level Hospital Facility in Greece: Systematic renew and retrospective analysis using R studio

**Name**
**Stathopoulou Kalliopi-Maria**

Master thesis

Name of Supervisor
Plagianakos Vasileios

Lamia 2019

## Περιεχόμενα

**Ευχαριστίες**

Η εκπόνηση της διπλωματικής εργασίας δεν θα μπορούσε να πραγματοποιηθεί χωρίς την συγκατάθεση και την βοήθεια από τον επιβλέπων καθηγητή μου, κ. Πλαγιανάκο Βασίλειο, γι' αυτό τον ευχαριστώ θερμά, για την τιμή που μου έκανε να συνεργαστώ μαζί του, την υποστήριξη που μου παρείχε καθ' όλη την διάρκεια του Μεταπτυχιακού Προγράμματος καθώς και για την εμπιστοσύνη που μου έδειξε ως επιβλέπων κατά την επιλογή του θέματος και την συγγραφή της Διπλωματικής εργασίας. Η αγαστή συνεργασία των φοιτητών με τους εκπαιδευτές είναι το βασικότερο κομμάτι της εκπαιδευτικής διαδικασίας και ο κ. Πλαγιανάκος αποτελεί λαμπρό παράδειγμα ακαδημαϊκού με ήθος, απλότητα και αδιαμφισβήτητα βαθιά γνώση του αντικειμένου.

Θα ήθελα ακόμη να ευχαριστήσω τον πατέρα μου για όλες τις θυσίες που έχει κάνει ώστε να έχω φτάσει εδώ που έχω φτάσει, όποιο σημείο και αν είναι αυτό. Του αφιερώνω αυτήν την διπλωματική εργασία ελπίζοντας να είναι περήφανος για εμένα.

Επίσης ένα μεγάλο ευχαριστώ στους καθηγητές του Μεταπτυχιακού Προγράμματος του Τμήματος Πληροφορικής και Υπολογιστικής Βιοϊατρικής της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας για τις πολύτιμες γνώσεις που μου παρείχαν.

Τέλος, δεν θα ήθελα να εξαιρέσω από το κομμάτι των ευχαριστιών μου το προσωπικό του Στρατιωτικού Νοσοκομείου Αθηνών για τα δεδομένα που μου έδωσαν και για την εμπιστοσύνη που μου έδειξαν, διότι συνέβαλλαν και αυτοί με τον τρόπο τους στην διεξαγωγή της έρευνας.

**Περίληψη**

Ο καρκίνος της κύστης των ωοθηκών συνεχίζει ακόμη και σήμερα να αποτελεί παράγοντα απειλής της ζωής για ένα μεγάλο αριθμό γυναικών παγκοσμίως.
Μια κύστη ωοθηκών είναι ένας σάκος ή θύλακας γεμάτος με υγρό ή άλλο ιστό που σχηματίζεται μέσα ή πάνω σε μια ωοθήκη. Οι κύστες των ωοθηκών είναι πολύ συχνές. Μπορούν να εμφανιστούν κατά τη διάρκεια των αναπαραγωγικών ετών ή μετά την εμμηνόπαυση. Οι κύστες των ωοθηκών μπορεί να είναι καλοήθεις (όχι καρκίνοι), οι οποίες απομακρύνονται μόνες τους χωρίς θεραπεία. Αλλά, μια κύστη μπορεί να είναι κακοήθης (καρκίνος).

**Σκοπός** της διπλωματικής εργασίας είναι η συσχέτιση και η στατιστική μελέτη των δημογραφικών δεδομένων και των συμπτωμάτων των ασθενών με την ιστολογική έκβαση των ωοθηκικών κύστεων (καλοήθεια ή κακοήθεια).

**Μέθοδος:** Για την διεξαγωγή της έρευνας χρησιμοποιήθηκαν δεδομένα από την βάση δεδομένων του Στρατιωτικού Νοσοκομείου Αθηνών του Γυναικολογικού Τμήματος. Η ανίχνευση των δεδομένων έγινε με τη χρήση περιγραφικής στατιστικής και της συσχέτισης των κατηγορικών μεταβλητών με τους στατιστικούς ελέγχους και τη μηχανική μάθηση, χρησιμοποιώντας Λογιστική Παλινδρόμηση, Δέντρα Ταξινόμησης, Random Forest καθώς και τους αλγορίθμους SVM και Naïve Bayes με χρήση του R studio.

**Αποτελέσματα:** Στην έρευνα συμμετείχαν 480 γυναίκες, στις οποίες είχε εμφανιστεί κύστη ωοθήκης. Η κύστη σε κάθε γυναίκα ήταν είτε καλοήθης είτε κακοήθης. Οι παράγοντες που ήταν στατιστικά σημαντικοί στην εμφάνιση κακοήθειας παρατηρήθηκαν στις γυναίκες που βρίσκονταν στην εμμηνόπαυση, είχαν τουλάχιστον ένα παιδί, είχαν προσωπικό ιστορικό καρκίνου του μαστού και σε συγκεκριμένα συμπτώματα που παρουσίασαν και ήταν ο λόγος εισαγωγής τους στο νοσοκομείο. Αυτά είναι το κοιλιακό φούσκωμα, οι ανωμαλίες της εμμήνου ρύσεως και το αίσθημα διάχυτου κοιλιακού πόνου.

**ABSTRACT**

The Cancer of ovarian cyst continues still be a life-threatening factor for a large number of women in the world.
An ovarian cyst is a sac or pill filled with fluid or other tissue that is formed in or on an ovary. Ovarian cysts are very common. They can appear during reproductive years or after menopause. Ovarian cysts can be benign (not cancerous), which are removed alone without treatment. But, a cyst can be malignant (cancer).

**The aim** of the thesis is the correlation and the statistical study of the demographic data and the symptoms of the patients with the histological outcome of the ovarian cysts (benign or malignant).

**Method:** Data from database of the Military Hospital of Athens of Gynecology Department was used to conduct the survey. The detection of data was done using descriptive statistics and the correlation of categorical variables with statistical checks and machine learning, using Logistic Regression, Classification Trees, Random Forest and algorithms (SVM, Naïve Bayes) using R studio.

**Results:** In the survey participated 480 women who had ovarian cyst. The cyst in each woman was either benign or malignant. The factors that were statistically significant in the occurrence of malignancy were observed in menopausal women, had at least one child, had a history of breast cancer and specific symptoms that they presented and was the reason for their admission to the hospital. These are abdominal bloating, menstrual abnormalities and the feeling of diffuse abdominal pain.

# Θεωρητικό μέρος

# 1.Εισαγωγή

## 1.1 Εισαγωγή στις Κύστεις Ωοθηκών

*Α. ΑΝΑΤΟΜΙΑ*

### Α.1. ΚΟΛΠΟΣ

Ο κολεός (κόλπος) είναι ένας ινομυώδης σωλήνας ο οποίος συνδέει την μήτρα με την σχισμή του αιδοίου. Στο άνω τμήμα του κόλπου βρίσκεται ο τράχηλος της μήτρας. Ο κόλπος αποτελεί το κύριο όργανο της συνουσίας καθώς είναι το όργανο εκείνο που υποδέχεται το πέος κατά την διάρκεια της σεξουαλικής πράξης. Χρησιμεύει επίσης, για την έξοδο του εμβρύου από την μήτρα κατά την διάρκεια του τοκετού καθώς και την αποβολή του εμμηνορυσιακού υλικού στο τέλος του κύκλου.

Ο κόλπος βρίσκεται ανάμεσα στην ουρήθρα και τον πυθμένα της ουροδόχου κύστης προς τα εμπρός και το ορθό προς τα πίσω. Η πορεία του κόλπου είναι τέτοια ώστε ο άξονάς του σχεδόν σχηματίζει γωνία $45^{o}$ με το επίπεδο του περινέου. Η γωνία που σχηματίζει ο άξονας του κόλπου με τον άξονα της μήτρας ποικίλλει (πρόσθια κλίση κάμψη, ευθειωρία, οπίσθια κλίση κάμψη). Το μήκος του κόλπου είναι περίπου 7,5 εκ. κατά το πρόσθιο τοίχωμά του και 9 εκ. κατά το οπίσθιο, δηλαδή το οπίσθιο τοίχωμα προσφύεται σε υψηλότερο σημείο από το πρόσθιο.

Το σύνολο των ελαστικών και κολλαγόνων ινών του κόλπου, του προσδίδει μία ιδιαίτερη ικανότητα αύξησης της διαμέτρου του κατά την διάρκεια του τοκετού, ώστε να μπορέσει να εξέλθει το έμβρυο. Το σπέρμα αθροίζεται στον οπίσθιο κολπικό θόλο κατά την συνουσία και στη συνέχεια με την βοήθεια της τραχηλικής βλέννης

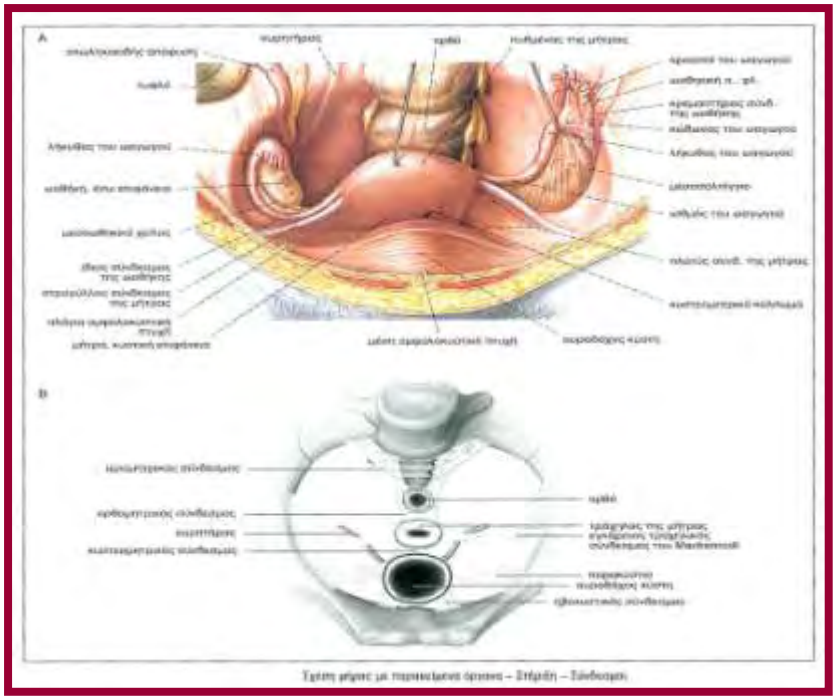
προστατεύεται και διευκολύνεται η άνοδος των σπερματοζωαρίων στην μήτρα. Κατά την γεροντική ηλικία, ο κόλπος συρρικνώνεται και στους δύο άξονες και οι θόλοι του κόλπου εξαφανίζονται.

Το τοίχωμα του κόλπου έχει πάχος περίπου 1,5 εκ. και αποτελείται από μέσα προς τα έξω από τον βλεννογόνο, τον μυϊκό και τον ινώδη χιτώνα. Ο τελευταίος αποτελείται από πυκνό συνδετικό ιστό. Ο μυϊκός χιτώνας αποτελείται από λείες μυϊκές ίνες που σχηματίζουν δύο στοιβάδες, την έξω και την έσω. Ο βλεννογόνος του κόλπου συμφύεται στερεά με τον μυϊκό χιτώνα. Κατά την ήβη το επιθήλιο του χορίου εμπλουτίζεται σε γλυκογόνο και με την βοήθεια των γαλακτοβάκιλλων του Doderlein (φυσιολογική χλωρίδα του κόλπου), διασπάται σε γαλακτικό οξύ με αποτέλεσμα το pH του κόλπου να παραμένει όξινο και να προστατεύεται από τον αποικισμό άλλων παθογόνων μικροοργανισμών.

Ο κόλπος αιματώνεται από τους κατιόντες κλάδους της μητριαίας αρτηρίας, από την κάτω κυστική, την έσω αιδοιική και την μέση αιμορροιδική. Το φλεβικό δίκτυο του κόλπου αποτελείται από το μητροκολεικό φλεβώδες πλέγμα που καταλήγει στην έσω λαγόνιο φλέβα. Τα λεμφογάγγλια εκβάλλουν στα έξω λαγόνια και επιπολής βουβωνικά λεμφογάγγλια[1].

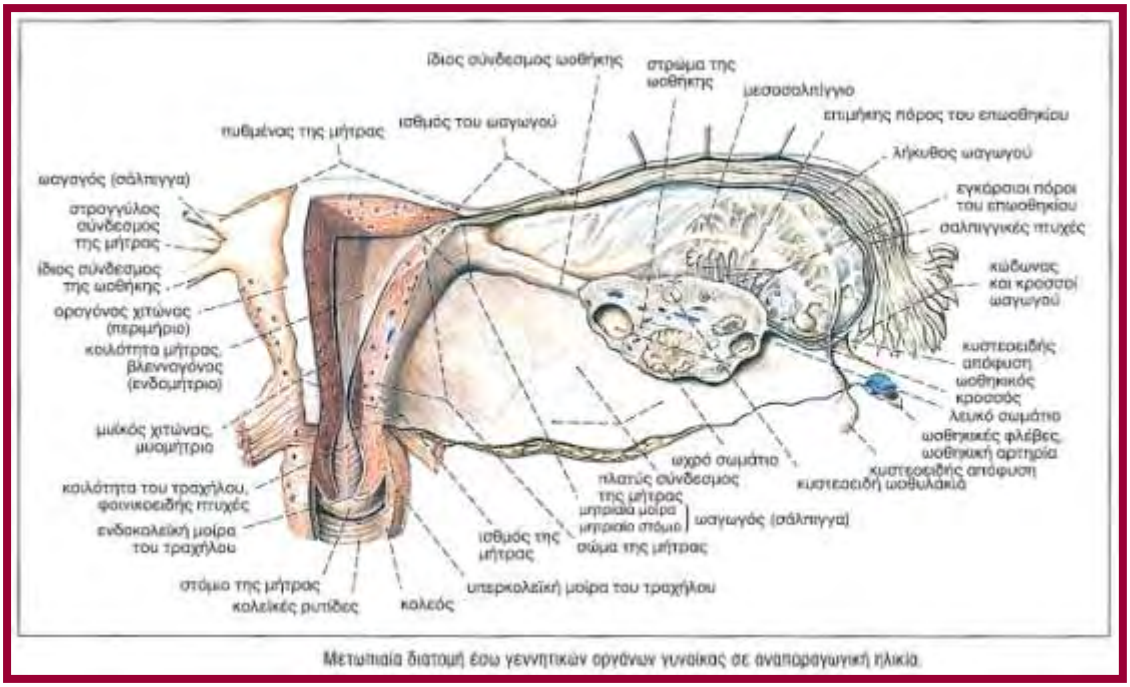
## Α.2. ΜΗΤΡΑ

Η μήτρα είναι ένα κοίλο μυώδες όργανο απιοειδούς σχήματος που βρίσκεται στην ελάσσονα πύελο και το οποίο συνδέει τον κόλπο με τους ωαγωγούς (σάλπιγγες). Διαθέτει τρεις χιτώνες (στοιβάδες) συνολικού πάχους 10-20 mm, οι οποίες από την κοιλότητα προς την επιφάνεια είναι οι εξής: βλεννογόνος (ενδομήτριο), μυϊκός χιτώνας (μυομήτριο), ορογόνος χιτώνας (περιμήτριο). Η μήτρα είναι το κύριο όργανο της κύησης καθώς είναι αυτή που θα υποδεχτεί το γονιμοποιημένο ωάριο στο στάδιο της βλαστοκύστης και θα αποτελέσει τον χώρο όπου θα αναπτυχθεί το κύημα. (Εικ. 1)



***Εικόνα 1:***
*Α) σχέση της μήτρας με τα λοιπά όργανα της ελάσσονος πυέλου*
*Β) σύνδεσμοι της μήτρας (Περιγραφική Ανατομική Άγιος Α.-Τα σπλαχνα)*

Η μήτρα, όπως προείπαμε, βρίσκεται τοπογραφικά εντός της ελάσσονος πυέλου ,πίσω από την ουροδόχο κύστη και μπροστά από το ορθό (απευθυσμένο) (Εικ. 1). Σε μετωπιαία διατομή η μήτρα έχει τριγωνικό περίπου σχήμα, με την βάση του τριγώνου να στρέφεται προς τα πάνω και την κορυφή προς τα κάτω (Εικ. 2). Το μήκος της είναι περίπου 8 εκ., το πλάτος της 5 εκ., ενώ το βάρος της σε μία γυναίκα αναπαραγωγικής ηλικίας περίπου 50 γραμμάρια.




***Εικόνα 2****:μετωπιαία διατομή μήτρας και λοιπών έσω γεννητικών οργάνων(Περιγραφική Ανατομική Άγιος Α.-Τα σπλαχνα)*

Η μήτρα διακρίνεται στον πυθμένα, στο σώμα και στον τράχηλο της μήτρας. Όριο ανάμεσα στον πυθμένα και το σώμα αποτελεί η μεσοσαλπιγγική γραμμή, μία νοητή γραμμή που συνδέει τα σημεία στα οποία εισέρχονται οι ωαγωγοί στην μήτρα, ενώ όριο ανάμεσα στο σώμα και τον τράχηλο αποτελεί ο ισθμός που αποτελεί και το στενότερο τμήμα της μήτρας.Από τον πυθμένα της μήτρας εκπορεύονται ο στρογγύλος σύνδεσμος της μήτρας (μπροστά), ο ωαγωγός (στο μέσο) και ο ίδιος σύνδεσμος της ωοθήκης (πίσω).

Το σώμα της μήτρας παρουσιάζει δύο επιφάνειες, την πρόσθια και την οπίσθια, καθώς και δύο πλάγια χείλη. Η πρόσθια επιφάνεια έρχεται σε επαφή με την ουροδόχο κύστη μέσω της κυστεομητρικής πτυχής που αποτελεί τμήμα του περιτοναίου, ενώ η οπίσθια επιφάνεια επαλείφεται από την ορθομητρική πτυχή που είναι επίσης τμήμα του περιτοναίου και έρχεται σε επαφή με το σιγμοειδές και το ορθό που είναι τα τελικά τμήματα του παχέος εντέρου. Στα πλάγια χείλη της μήτρας προσφύονται οι πλατείς σύνδεσμοι καθώς και τα μητριαία αγγεία. Οι πλατείς σύνδεσμοι είναι πτυχές του περιτοναίου που αποτελούνται από δυο πέταλα (πρόσθιο και οπίσθιο) και εκτείνονται μεταξύ της μήτρας και των πλάγιων πυελικών τοιχωμάτων. Ανάμεσα στα δυο αυτά πέταλα των πλατέων συνδέσμων υπάρχει άφθονος χαλαρός συνδετικός ιστός.

Ο τράχηλος της μήτρας αποτελεί και το κατώτερο τμήμα της και συνδέει την μήτρα με τον κολεό (κόλπο). Ο άξονας του τραχήλου και της μήτρας, συνήθως, δεν συμπίπτουν καθώς σχηματίζεται μία γωνία μεταξύ τους (κάμψη της μήτρας), ενώ στις σπάνιες περιπτώσεις που συμπίπτουν έχουμε την λεγόμενη «ευθειωρία» της μήτρας. Στα πλάγια του τραχήλου βρίσκονται τα παραμήτρια, που αποτελούν το έδαφος των πλατέων συνδέσμων. Μέσα στα παραμήτρια παρατηρείται η πορεία της μητριαίας αρτηρίας προς την μήτρα καθώς και η πορεία του ουρητήρα προς την ουροδόχο κύστη. Οι δύο προηγούμενες δομές χιάζονται σε απόσταση 2 εκατοστών από τα πλάγια του τραχήλου και το σημείο αυτό αποτελεί σημαντικό χειρουργικό σημείο στις περισσότερες γυναικολογικές επεμβάσεις.

Η κοιλότητα της μήτρας, σε σχέση με τον συνολικό όγκο του συγκεκριμένου οργάνου, είναι μικρή και αυτό οφείλεται στα παχιά τοιχώματά του. Η κοιλότητα της μήτρας επαλείφεται από το ενδομήτριο (βλεννογόνο) που αποτελείται από χόριο και αδένες. Το ενδομήτριο ακολουθεί τις μηνιαίες ορμονικές μεταβολές όπως αυτές καθορίζονται από τις ωοθήκες και τελικά τμήμα αυτού αποπίπτει με την μορφή της εμμήνου ρύσεως.

Ο μυϊκός χιτώνας (μυομήτριο), αποτελεί και το μεγαλύτερο τμήμα της μήτρας. Αποτελείται από λείες μυϊκές ίνες, αιμοφόρα και λεμφοφόρα αγγεία καθώς και νεύρα. Οι μυϊκές ίνες διαπλέκονται σε διάφορες κατευθύνσεις σχηματίζοντας την έξω, μέση και έσω μυϊκή στοιβάδα. Η διάταξη των μυϊκών ινών έχει ιδιαίτερη σημασία κατά την φάση εξώθησης στον τοκετό καθώς επίσης και στον στραγγαλισμό των ακτινοειδών αρτηριακών κλάδων και την πρόληψη της αιμορραγίας μετά την έξοδο του κυήματος από την μήτρα.

Ο ορογόνος χιτώνας (περιμήτριο), αποτελεί τμήμα του περιτοναίου που καλύπτει ολόκληρη την οπίσθια επιφάνειά της καθώς και την υπερκολεϊκή μοίρα του τραχήλου. Ο ορογόνος δεν καλύπτει τα πλάγια χείλη της μήτρας καθώς και την πρόσθια υπερκολεϊκή μοίρα του τραχήλου.

Η μήτρα αγγειώνεται κυρίως από την μητριαία αρτηρία που είναι κλάδος της έσω λαγονίου αρτηρίας, ενώ παράλληλα δέχεται αγγείωση από ωοθηκική, την κολεϊκή και την κάτω επιγάστρια αρτηρία (Εικ. 3). Το φλεβικό δίκτυο της μήτρας εκβάλλει στο μητροκολεικό και αιμορροιδικό φλεβικό πλέγμα που καταλήγουν με την σειρά τους στην έσω λαγόνια και έσω σπερματική φλέβα. Η λεμφική άρδευση γίνεται στα αορτικά, τα έξω λαγόνια, στα επίπολής βουβωνικά, στα έσω λαγόνια καθώς και στα κοινά λαγόνια, αιμορροιδικά και ιερά λεμφογάγγλια.

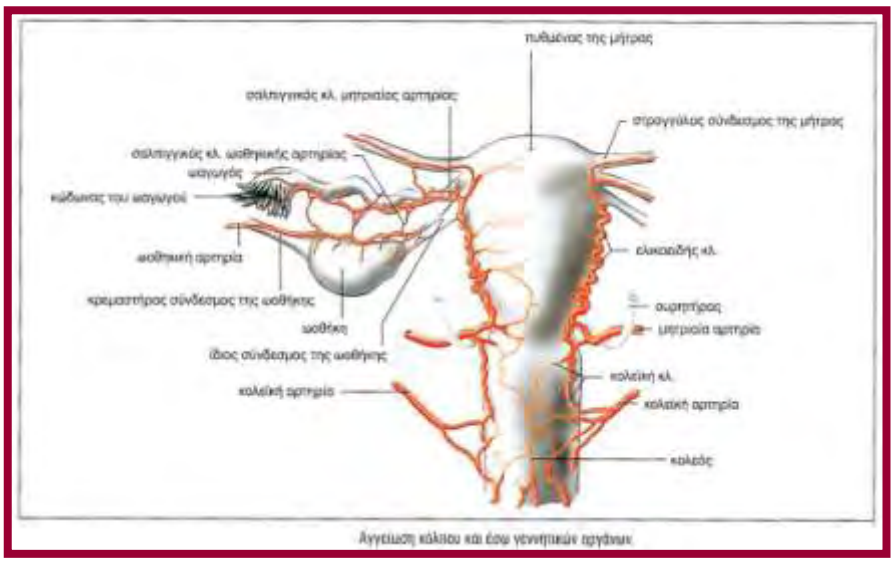
Η μήτρα μεταβάλλεται σε διαστάσεις, θέση και σύσταση ανάλογα με την ηλικία της γυναίκας και το ορμονολογικό της προφίλ. Κατά την νεογνική περίοδο η μήτρα δεν βρίσκεται στην ελάσσονα πύελο αλλά σε υψηλότερη θέση. Στην ενήλικη γυναίκα η μήτρα βρίσκεται στην τυπική θέση που περιγράφηκε ενώ κατά την γεροντική ηλικία η μήτρα γίνεται πλέον ατροφική. Κατά την διάρκεια της εμμηνορρυσίας η μήτρα αυξάνει σε μέγεθος και έχει μαλακότερη υφή. Στην διάρκεια της εγκυμοσύνης η μήτρα αποκτά το μεγαλύτερο δυνατό μέγεθος λόγω της διάτασης της από το κύημα, το αμνιακό υγρό που το περιβάλλει καθώς και την υπερτροφία των μυϊκών ινών της. Κατά την περίοδο της λοχείας, αρχίζει μία σταδιακή παλινδρόμηση της μήτρας στα προ-εγκυμοσύνης επίπεδα, η οποία ολοκληρώνεται περίπου 8 εβδομάδες μετά τον τοκετό[2-4].

## Α.3. ΣΑΛΠΙΓΓΑ

Οι ωαγωγοί, ή σάλπιγγες είναι δύο μυώδεις σωλήνες μήκους 10-12 cm, οι οποίοι προσεκβάλλουν από την κοιλότητα της μήτρας στην περιοχή του πυθμένα της και αφού πορευθούν στο άνω χείλος του πλατέως συνδέσμου ακουμπούν με το ελεύθερο άκρο τους την σύστοιχη ωοθήκη (Εικ. 2). Οι σάλπιγγες αποτελούν τα ομόλογα όργανα των σπερματικών πόρων του άνδρα. Διαφέρουν από αυτούς ανατομικά (δεν εισδύουν στην ωοθήκη), αλλά και λειτουργικά, αφού δεν μεταφέρουν μόνο το ωάριο αλλά επιπλέον διευκολύνουν, διαμέσου ρεύματος υγρών, την μετακίνηση των σπερματοζωαρίων και την ανάμιξη τους με το ωάριο ώστε να επιτευχθεί η γονιμοποίηση του τελευταίου. Από περιγραφική άποψη διακρίνονται στον ωαγωγό από έσω προς τα έξω η μητριαία μοίρα, ο ισθμός, η λήκυθος (εκεί συμβαίνει συνήθως η γονιμοποίηση του ωαρίου και διευκολύνεται από την έντονη πτύχωση βλεννογόνου) και ο κώδωνας με τους κροσσούς. Το τοίχωμα του ωαγωγού συνίσταται από έξω προς τα μέσα από τον ορογόνο χιτώνα, που είναι τμήμα του περιτοναίου και περιβάλλει εξ’ ολοκλήρου τον ωαγωγό, το μυϊκό χιτώνα με δύο στιβάδες λείων μυϊκών ινών, μία έξω επιμήκη και μία έσω κυκλοτερή και τον βλεννογόνο χιτώνα που αποτελείται από επιθήλιο με έναν στοίχο κυλινδρικών κροσσωτών κύτταρων και διάσπαρτα εκκριτικά κύτταρα.Τα τελευταία επικάθονται πάνω σε έναν ατελή βασικό υμένα και χόριο στο οποίο περιέχονται αιμοφόρα και λεμφοφόρα αγγεία και νεύρα. Οι αρτηρίες που αιματώνουν τον ωαγωγό είναι κλάδοι

της ωοθηκικής και της μητριαίας αρτηρίας, ενώ η φλεβική αποχέτευση γίνεται με την ωοθηκική, την μητριαία και την κάτω επιγάστρια φλέβα[5] (Εικ. 3).

Η δόνηση των κροσσωτών κυττάρων δημιουργεί ρεύμα στο υγρό που περιέχεται στον αυλό της σάλπιγγας, το οποίο κατευθύνεται προς την μήτρα. Η σημασία των κροσσωτών κυττάρων του επιθηλίου του ωαγωγού στην φυσιολογία της γονιμοποίησης είναι τόσο μεγάλη, ώστε από πολλούς υποστηρίζεται ότι καταστροφή 50% περίπου αυτών των κυττάρων καθιστά αδύνατη την γονιμοποίηση του ωαρίου έστω και αν ο αυλός της σάλπιγγας ελέγχεται διαβατός από υγρό η αέρα που διοχετεύεται σε αυτόν. Εκτός από την δόνηση των κροσσών στον μηχανισμό διαμόρφωσης αυτού του ρεύματος συμμετέχουν οι περισταλτικές κινήσεις του ωαγωγού καθώς και η δυνατότητα παραγωγής του υγρού με διαφορετικό ρυθμό από τις διάφορες μοίρες του ωαγωγού, ιδιαίτερα κατά το χρονικό διάστημα της πιθανής γονιμοποίησης. Το ρεύμα αυτό διευκολύνει την μεταφορά του ωαρίου, ενώ παρεμποδίζει την κίνηση των σπερματοζωαρίων, τα οποία όμως το υπερνικούν γιατί εμφανίζουν δική τους κινητικότητα που καθορίζεται από θετικό ρεοτακτισμό. Το λαβυρινθώδες της όψης του αυλού του ωαγωγού ευνοεί την ανάπτυξη συμφύσεων ύστερα από φλεγμονή (σαλπιγγίτιδα), η οποία έχει ως συνέπεια την μερική ή πλήρη απόφραξη του αυλού με αποτέλεσμα υπογονιμότητα της γυναίκας[6].




***Εικόνα 3:*** *αγγείωση έσωγεννητικών οργάνων θήλεως(Περιγραφική Ανατομική Άγιος Α.- Τα σπλαχνα)*

## Α.4. ΩΟΘΗΚΗ

Οι δύο ωοθήκες αποτελούν ομόλογα όργανα των όρχεων του άνδρα και θεωρούνται μικτοί αδένες καθώς παράγονται τόσο τα γεννητικά κύτταρα του θήλεος (ωάρια), όσο και ένα μεγάλο μέρος των γεννητικών ορμονών.

Οι ωοθήκες, από την διάπλαση του θήλεος εμβρύου ενδομητρίως και μέχρι και πριν την εφηβεία, βρίσκονται ψηλά στην πύελο περίπου στο ύψος της οσφυϊκής μοίρας της σπονδυλικής στήλης. Με την έναρξη της ήβης, οι ωοθήκες καταλαμβάνουν την τυπική τους θέση μέσα στους αντίστοιχους ωοθηκικούς βόθρους στα πλάγια της μήτρας, βρίσκονται δε αναρτημένοι στο οπίσθιο πέταλο του πλατέως συνδέσμου της μήτρας με την βοήθεια μιας περιτοναϊκής πτυχής που ονομάζεται μεσωοθήκιο.

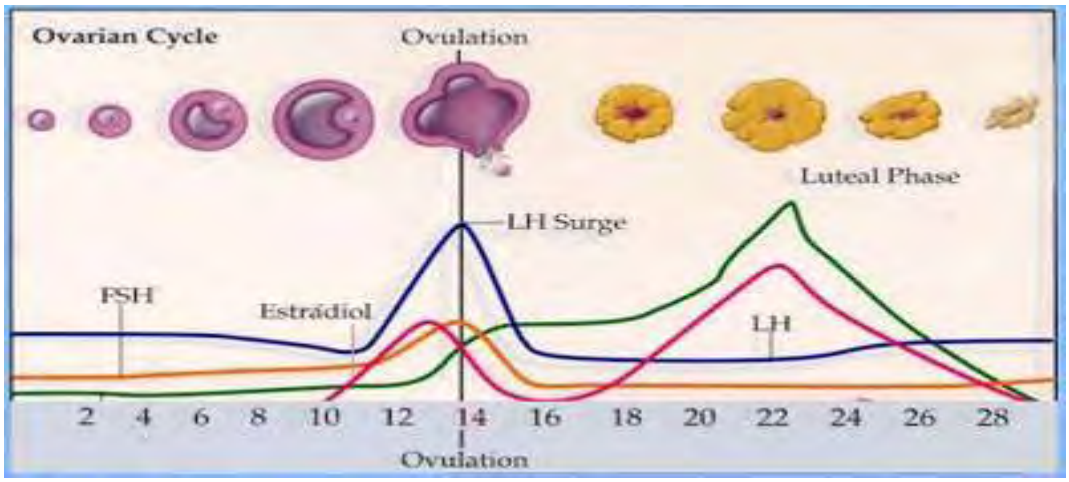
Οι ωοθήκες έχουν αμυγδαλοειδές σχήμα με περίπου 3 εκ. μήκος, 2 εκ. πλάτος και 1 εκ. πάχος. Το βάρος μίας φυσιολογικής ωοθήκης σε γυναίκα αναπαραγωγικής ηλικίας ανέρχεται στα 8 περίπου γραμμάρια. Η έσω επιφάνεια της ωοθήκης καλύπτεται κυρίως από τους κροσσούς του κώδωνα της σύστοιχης σάλπιγγας.

Η ωοθήκη αποτελείται από τρεις στοιβάδες οι οποίες από έξω προς τα έσω είναι το βλαστικό επιθήλιο, η φλοιώδης και η μυελώδης μοίρα. Στην ωοθήκη γίνεται η ωρίμανση των άωρων ωοθυλακίων προς ώριμα γεννητικά κύτταρα (ωάρια), διαμέσου συγκεκριμένης διαδικασίας ωρίμανσης.

Η ωοθήκη αιματώνεται από την ωοθηκική αρτηρία, η οποία εν συνεχεία αναστομώνεται με τον ωοθηκικό κλάδο της μητριαίας αρτηρίας. Το φλεβικό δίκτυο αρχίζει με τις ωοθηκικές φλέβες οι οποίες εν συνεχεία σχηματίζουν το ωοθηκικό φλεβώδες πλέγμα. Η λεμφική αποχέτευση γίνεται στα προαορτικά και παραορτικά λεμφογάγγλια, ενώ μέσω των στρογγύλων συνδέσμων της μήτρας και στα βουβωνικά λεμφογάγγλια[6]

*Β. ΦΥΣΙΟΛΟΓΙΑ ΕΜΜΗΝΟΡΡΥΣΙΑΚΟΥ ΚΥΚΛΟΥ*

Ο εμμηνορρυσιακός κύκλος της γυναίκας χωρίζεται σε 2 φάσεις, την ωοθυλακική και την ωχρινική. Το γεγονός που διαχωρίζει τις φάσεις αυτές είναι η ωοθυλακιορρηξία.




***Εικόνα 4****:Σχηματική απεικόνιση της σχέσης των υποφυσιακών ορμονών FSH-LH με τις ωοθηκικές ορμόνες οιστραδιόλη-προγεστερόνη στην ωοθυλακική και ωχρινική φάση ενός φυσιολοφικού κύκλου 28 ημερών (http://www.ovulation-calculator.com/menstrual-cycle.htm)*

Η διαδικασία των γεγονότων, κάθε στιγμή κατά την διάρκεια του εμμηνορρυσιακού κύκλου, εξαρτάται κυρίως από την παρουσία συγκεκριμένων ορμονών και τα επίπεδά τους.

Κατά την ωοθυλακική φάση του κύκλου, η οποία αρχίζει με την πρώτη μέρας της εμμήνου ρύσεως, συμβαίνει η επιλογή ενός μόνο ωοθυλακίου μέσα από μια δεξαμενή πολλών ωοθυλακίων, το οποίο θα ωριμάσει και θα οδηγηθεί στην ωορρηξία. Στην αρχή της ωοθυλακικής φάσης παρατηρείται μία αύξηση στα επίπεδα της FSH, η οποία αποδίδεται στην πτώση των επιπέδων προγεστερόνης και οιστραδιόλης του προηγούμενου κύκλου και άρση της ανασταλτικής τους ιδιότητας στην FSH. Με την έναρξη κάθε εμμηνορρυσιακού κύκλου διεγείρονται προς ανάπτυξη περίπου 15-20 ωοθυλάκια[7]. Η διέγερση αυτή γίνεται μέσω της FSH, η οποία διεγείρει έμμεσα την έκκριση οιστραδιόλης μέσω της αρωματοποίησης των ανδρογόνων που εκκρίνονται από την έξω θήκη του ωοθυλακίου[8]. Η FSH διεγείρει περαιτέρω την δημιουργία νέων υποδοχέων FSH, αυξάνοντας τελικά τα επίπεδα της οιστραδιόλης, η οποία παλίνδρομα αναστέλλει την αρχική έκκριση της FSH της οποίας τα επίπεδα τελικά φθίνουν[9].

Κάτω από φυσιολογικές συνθήκες, ένα μόνο ωοθυλάκιο είναι αυτό το οποίο θα οδηγηθεί σε ωοθυλάκιορρηξία ως "κυρίαρχο" ωοθυλάκιο. Τα υπόλοιπα ωοθυλάκια θα οδηγηθούν σε ατρησία και κυτταρικό θάνατο. Η διαδικασία με την οποία επιλέγεται το κυρίαρχο ωοθυλάκιο δεν είναι ακριβώς γνωστή. Πιστεύεται, ότι το κυρίαρχο ωοθυλάκιο διαθέτει εξαρχής ένα μεγάλο αριθμό υποδοχέων FSH. Καθώς τα επίπεδα της FSH φθίνουν στο τέλος της ωοθυλακικής φάσης – όπως προείπαμε – τα υπό ωρίμανση ωοθυλάκια πρέπει να λειτουργήσουν με χαμηλά επίπεδα FSH. Με αυτόν τον τρόπο είναι κατανοητό ότι το ωοθυλάκιο εκείνο που θα είναι πιο προετοιμασμένο, δηλαδή "εξοπλισμένο" με περισσότερους υποδοχείς FSH, θα μπορέσει να επιβιώσει και να αναπτυχθεί. Το κυρίαρχο ωοθυλάκιο ωριμάζει και εκκρίνει αυξανόμενα ποσά οιστρογόνων, τα οποία φτάνουν το μέγιστο στο τέλος της ωοθυλακικής φάσης του κύκλου. Τα αυξημένα επίπεδα της οιστραδιόλης (>200 pg/ml για περισσότερο από 50 ώρες) προκαλούν, μέσω θετικής παλίνδρομης αλληλορύθμισης, την εκκριτική αιχμή της LH που είναι απαραίτητη για την ρήξη του ώριμου ωοθυλακίου[10].

Κάτω από την επίδραση της LH, το πρωτογενές ωοκύτταρο ολοκληρώνει την πρώτη μειωτική διαίρεση και διαιρείται σε δευτερογενές ωοκύτταρο και πρώτο πολικό σωμάτιο. Η εκκριτική αιχμή της LH προκαλεί απελευθέρωση πρωτεολυτικών ενζύμων, τα οποία προκαλούν με την σειρά τους διάβρωση του τοιχώματος του ωοθυλακίου, αγγειογένεση στο ωοθυλακικό τοίχωμα και απελευθέρωση προσταγλανδινών[11-12]. Μέσω αυτής της διαδικασίας επιτυγχάνεται ρήξη του τοιχώματος του ωοθυλακίου και απελευθέρωση του ωοκυττάρου στην σάλπιγγα.

Η ωχρινική φάση του κύκλου χαρακτηρίζεται από την ωχρινοποίηση του ωοθυλακίου που απέμεινε μετά την απελευθέρωση του ωαρίου. Τα κοκκώδη κύτταρα, τα κύτταρα της θήκης καθώς και ο περιβάλλων συνδετικός ιστός, μετατρέπονται στο ωχρό σωμάτιο. Το σημαντικότερο γεγονός στην φάση αυτή είναι η ενεργοποίηση των κοκκωδών κυττάρων από την προηγηθείσα αιχμή της LH για την παραγωγή προγεστερόνης, που είναι και η κύρια ορμόνη του δεύτερου μισού του εμμηνορρυσιακού κύκλου[13]. Το μέγιστο της έκκρισης της προγεστερόνης επιτυγχάνεται 8 ημέρες μετά την ωορρηξία, οπότε και το ωχρό σωμάτιο διαθέτει και την μέγιστη αγγείωση. Η μεγάλη συγκέντρωση της προγεστερόνης προκαλεί, μέσω αρνητικής παλίνδρομης αλληλορύθμισης, πτώση των επιπέδων της GnRH, με αποτέλεσμα πτώση και στις υποφυσιακές ορμόνες FSH και LH. Αυτό έχει ως αποτέλεσμα το ωχρό σωμάτιο να μην δέχεται την δράση των ορμονών και να οδηγείται σε ατρησία 14 ημέρες μετά την ωορρηξία οπότε και

μετατρέπεται στο λευκό σωμάτιο. Με την πτώση τόσο των οιστρογόνων όσο και της προγεστερόνης αίρεται η ανασταλτική δράση στην GnRH και έτσι παρατηρείται σταδιακή αύξηση στην FSH που οδηγεί στον επόμενο κύκλο[14].

*Γ. Κύστεις Ωοθηκών*

Οι ωοθήκες της γυναίκας μπορεί να υποστούν διάφορες μεταβολές κατά την διάρκεια της αναπαραγωγικής της ηλικίας καθώς και στην εμμηνόπαυση. Οι μεταβολές αυτές μπορεί να επιφέρουν την εμφάνιση κύστεων. Οι κύστεις είναι «σάκοι» με υγρό περιεχόμενο που είτε αποδρομούν από μόνες τους, είτε υφίστανται αυτόματη ρήξη εντός της ενδομητρίου κοιλότητας, είτε δημιουργούν προβλήματα στην καθημερινότητα της γυναίκας και πρέπει να αφαιρεθούν χειρουργικά.

Οι ωοθηκικές κύστεις συνήθως εμφανίζονται στην αναπαραγωγική ηλικία και η πλειονότητα αυτών είναι καλοήθεις. Γυναίκες που βρίσκονται ορμονικά στην εμμηνόπαυση εμφανίζουν μεγαλύτερο κίνδυνο για εμφάνιση κύστεων με κακοήθη χαρακτηριστικά.

## Γ.1. Καλοήθεις κύστεις

- **Λειτουργικές κύστεις**

Οι πιο συχνές ωοθηκικές κύστεις είναι οι λειτουργικές κύστεις. Υπάρχουν δύο τύποι λειτουργικών κύστεων, η ωοθυλακική κύστη και το κυστικό ωχρό σωμάτιο. Οι κύστεις αυτές συναντώνται κατά την αναπαραγωγική ηλικία της γυναίκας και το ιδιαίτερο χαρακτηριστικό τους είναι ότι εξαφανίζονται αυτόματα συνήθως μετά το πέρας 1-2 κύκλων [15].

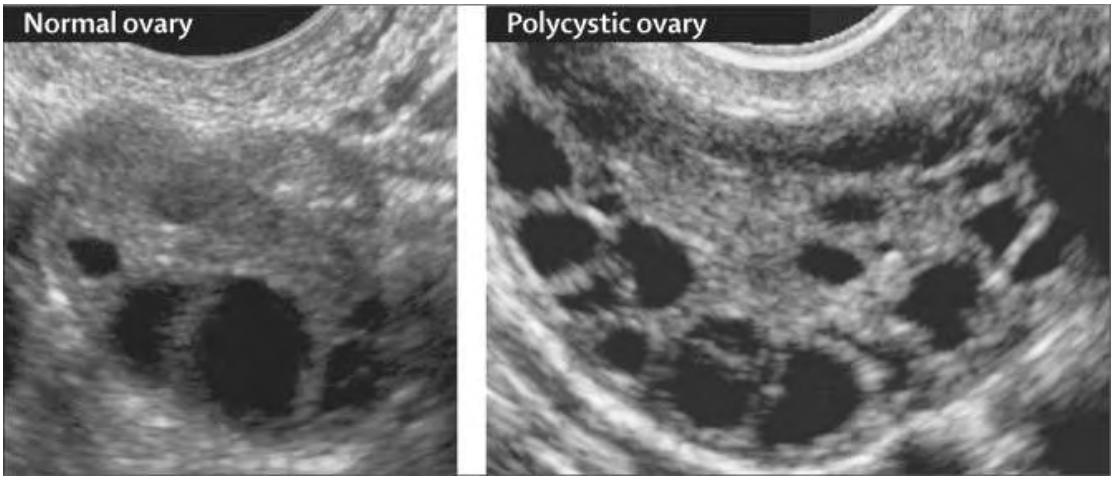
Οι ωοθυλακικές κύστεις είναι ασυμπτωματικές κύστεις οι οποίες συχνά είναι ψηλαφητές και μπορεί να φτάσουν το μέγεθος των 20-25cm.Αποτελούν ωοθυλάκια τα οποία απέτυχαν να υποστούν ρήξη κατά την ωορρηξία εξαιτίας παροδικών ενδοκρινολογικών διαταραχών και συνέχισαν να αυξάνονται σε μέγεθος.

Ο δεύτερος τύπος λειτουργικής κύστης είναι το κυστικό ωχρό σωμάτιο. Αυτό σχηματίζεται μετά την διαδικασία της ωορρηξίας. Τα κυστικά αυτά μορφώματα έχουν ιδιαίτερα αυξημένη περιφερική αγγείωση καθώς και αιμορραγικό περιεχόμενο. Το μέγεθός τους είναι συνήθως

μικρότερο από 2,5cm αλλά μερικές φορές τριπλασιάζονται σε μέγεθος εξαιτίας συνεχιζόμενης εσωτερικής αιμορραγίας(αιμορραγική κύστη).Όπως και οι ωοθυλακικές κύστεις έτσι και το κυστικό ωχρό παλινδρομεί αυτόματα και συνήθως δεν απαιτείται θεραπεία[16].

- **Σύνδρομο Πολυκυστικών Ωοθηκών**

Οι ασθενείς με σύνδρομο πολυκυστικών ωοθηκών έχουν ιδιαίτερα σύνθετο ιστορικό καθώς και υπερηχογραφικών χαρακτηριστικών. Στο σύνδρομο αυτό ,το οποίο είναι άγνωστης αιτιολογίας, οι ωοθήκες εμφανίζουν πολλά κυστικά μορφώματα περιφερικά στην ωοθήκη τα οποία όμως δεν ξεπερνούν το μέγεθος των 6-7mm. Η θεραπεία των κύστεων αυτή είναι φαρμακευτική και όχι χειρουργική[17].



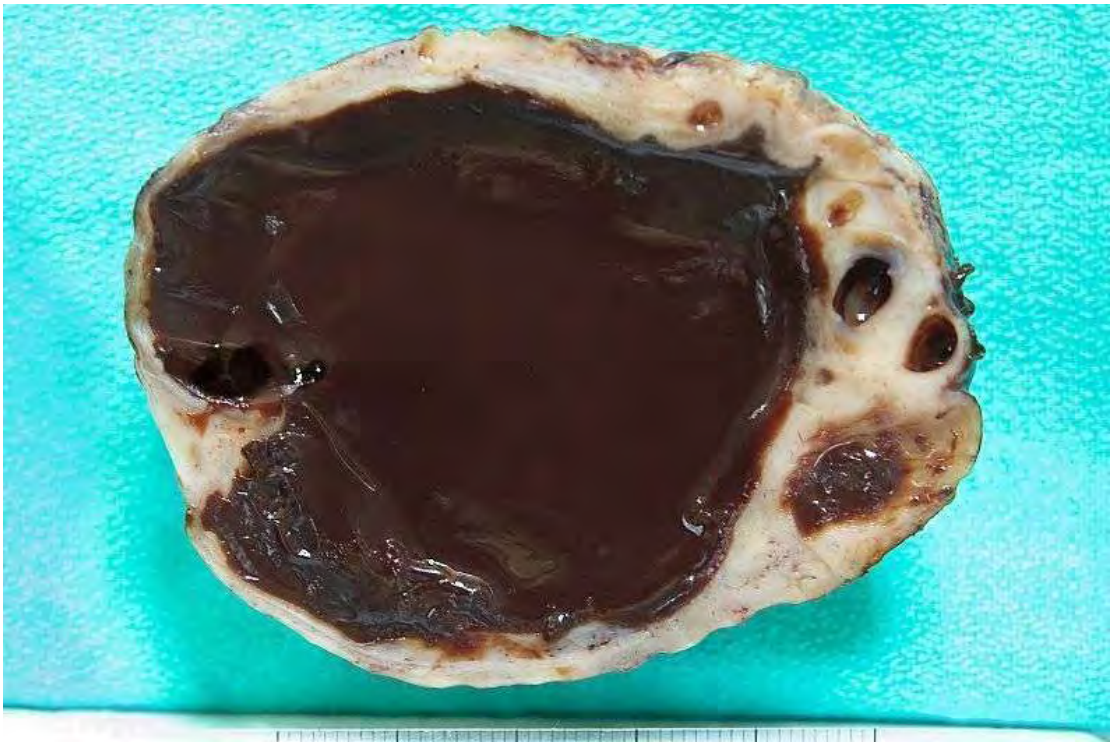

Εικονα : πολυκυστική ωοθήκη στο υπερηχογράφημα

- **Παρα-ωοθηκικές και παρα-σαλπιγγικές κύστεις**

Οι κύστες αυτές είναι εμβρυολογικής καταγωγής και βρίσκονται στους συνδέσμους δίπλα στην ωοθήκη. Ανευρίσκονται σε γυναίκες αναπαραγωγικής ηλικίας και είναι μικρές σε μέγεθος και απλές στην εμφάνιση[18].

- **Ενδομητριοειδής κύστη**

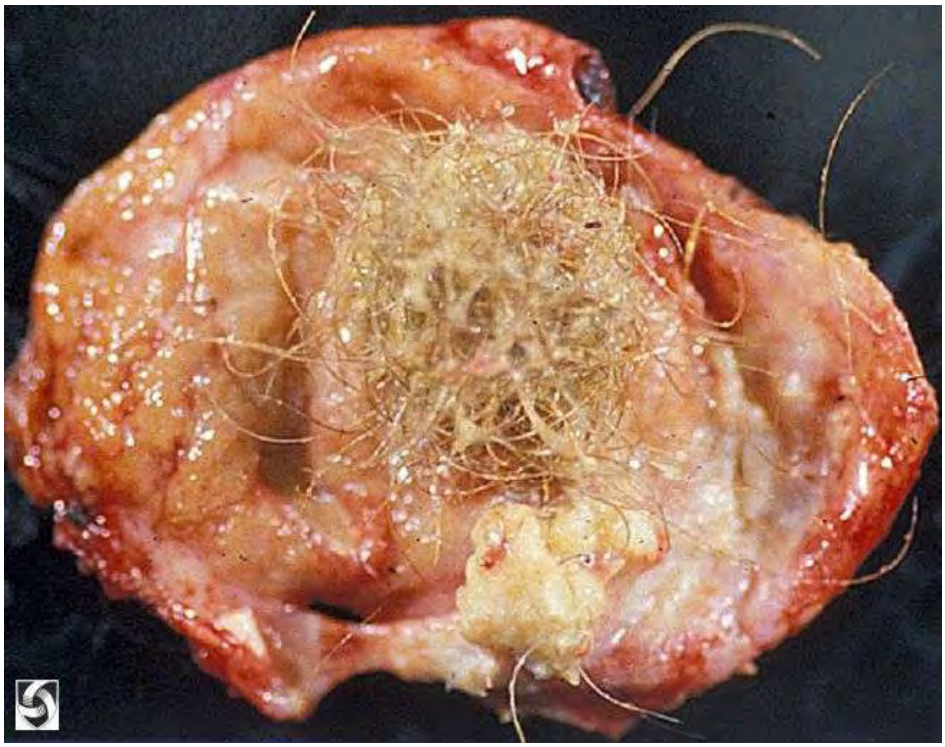
Οι κύστεις αυτές αποτελούν επιπλοκή μίας πάθησης γνωστής ως ενδομητρίωση κατά την οποία ενδομητρικός ιστός βρίσκεται σε άλλες απομακρυσμένες εκτός της φυσιολογικής του θέσης και ακολουθεί τις ορμονικές μεταβολές του εμμηνορυσιακού κύκλου. Οι συνεχόμενες κυκλικές αιμορραγίες εντός της κύστης αυτής δημιουργού τις ενδομητριωσικές ή σοκολατοειδείς κύστεις. Οι κύστεις αυτές προκαλούν δυσμηνόρροια (πόνο κατά την περίοδο), δυσπαρεύνια (πόνο κατά την επαφή), καθώς και υπογονιμότητα. Για τους προηγούμενους λόγους οι κύστεις αυτές αντιμετωπίζονται χειρουργικά[19].



Εικόνα : παρασκεύασμα ωοθήκης με ενδομητριωσική κύστη

- **Δερμοειδείς κύστεις**

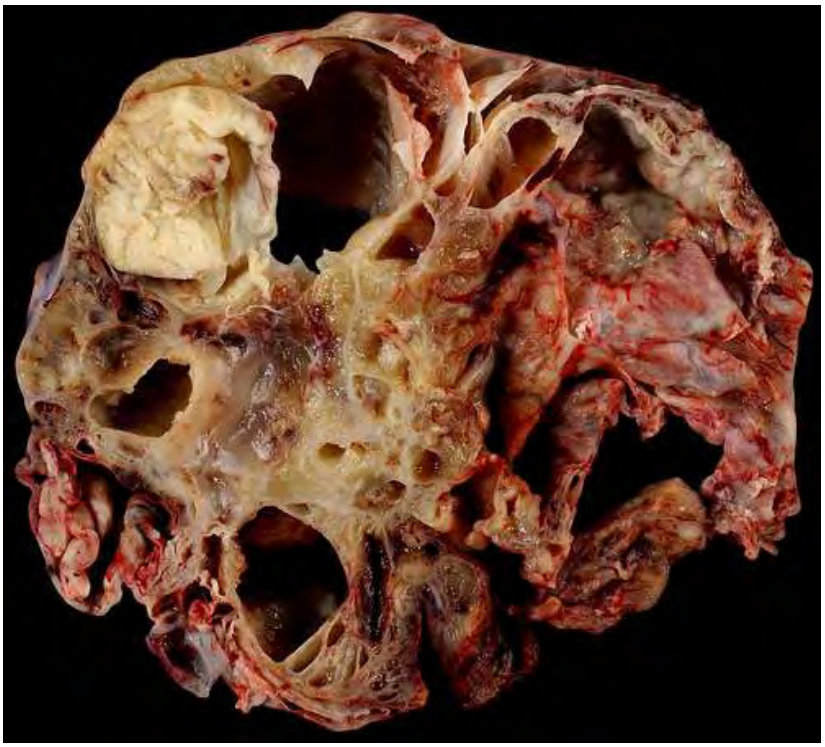
Η δερμοειδής κύστη είναι ένας καλοήθης όγκος εμβρυολογικής προέλευσης. Αποτελεί το 20% όλων των κύστεων των ωοθηκών και είναι ο πιο συχνός όγκος σε γυναίκες κάτω των 45 ετών. Οι δερμοειδείς κύστεις μπορεί να φτάσουν τα 10cm σε μέγεθος και περιέχουν εμβρυικά υπολείμματα όπως δόντια, τρίχες, νύχια και λίπος[20].



Εικόνα : παρασκεύασμα δερμοειδούς κύστης ωοθήκης

## Γ.2. Κακοήθεις κύστες

Το πρωταρχικό μέλημα κατά την ανεύρεση μία ωοθηκικής κύστης είναι ο αποκλεισμός της κακοήθειας. Ο ωοθηκικός καρκίνος είναι λιγότερο συχνός σε σχέση με τις άλλες κακοήθειες του γυναικείου αναπαραγωγικού συστήματος, αλλά και πιο καταστροφικός καθώς εμφανίζει συμπτώματα στα τελευταία στάδια της νόσου όπου οι θεραπευτικοί χειρισμοί είναι περιορισμένοι. Ο πιο συχνός τύπος είναι ο επιθηλιακός καρκίνος των ωοθηκών, ορώδης και βλεννώδης[21].



Εικόνα: βλεννώδες κυσταδενοκαρκίνωμα ωοθήκης

## 1.2 Εισαγωγικές Παρατηρήσεις

Η έννοια της μάθησης (Learning), της εξόρυξης δεδομένων (Data Mining) με τις αντίστοιχες σύγχρονες έννοιες της μηχανικής μάθησης (Machine Learning) και της στατιστικής μάθησης (Statistical Learning) συνδέονται και αλληλοεπιδρούν μεταξύ τους όσον αφορά θεωρητικά και εφαρμοσμένα προβλήματα των υπολογιστικών επιστημών (computing sciences) και των ποικίλων εφαρμογών τους.

Η "μάθηση" μπορεί να ορίζεται ως η "ικανότητα βελτίωσης της απόδοσης μετά την παρακολούθηση δεδομένων". Κατά τη διάρκεια των τελευταίων δύο δεκαετιών, σημειώθηκε έκρηξη εφαρμοσμένης και θεωρητικής εργασίας για την εκμάθηση μηχανών. Οι εφαρμογές των μεθόδων μάθησης είναι πανταχού παρούσες: περιλαμβάνουν συστήματα για ανίχνευση και αναγνώριση προσώπου, πρόβλεψη χρηματιστηριακών αγορών και πρότυπα καιρού, αναγνώριση ομιλίας, εκμάθηση προτιμήσεων αναζήτησης χρήστη, τοποθέτηση σχετικών διαφημίσεων και πολλά άλλα. Η επιτυχία αυτών των εφαρμογών ήταν παράλληλη με μια καλά αναπτυγμένη θεωρία. Θα το ονομάσουμε αυτό το τελευταίο κλάδο της μηχανικής μάθησης - «θεωρία μάθησης».

Πολλές εργασίες που θέλουμε να εκτελέσουν οι υπολογιστές δεν μπορούν να κωδικοποιηθούν. Τα προγράμματα πρέπει να προσαρμοστούν. Ο στόχος είναι τότε να κωδικοποιηθεί, για μια συγκεκριμένη εφαρμογή, όσο το δυνατόν περισσότερη γνώση σχετικά με το συγκεκριμένο τομέα και να αφήνεται αρκετή ευελιξία ώστε το σύστημα να βελτιώσει την παρακολούθηση των δεδομένων.

Είναι ευρέως αναγνωρισμένο ότι δεν υπάρχει κανένας αλγόριθμος ενιαίας μάθησης που θα λειτουργεί παγκοσμίως (θα κάνουμε αυτή τη δήλωση μαθηματικά ακριβής). Δεν είναι ο στόχος μας να κάνουμε τους υπολογιστές να μαθαίνουν τα πάντα ταυτόχρονα: κάθε εφαρμογή απαιτεί προηγούμενη γνώση από τον εμπειρογνώμονα. Ο στόχος της θεωρίας της μάθησης είναι να αναπτύξει γενικές κατευθυντήριες γραμμές και αλγόριθμους και να αποδείξει τις εγγυήσεις σχετικά με τις μαθησιακές επιδόσεις υπό διάφορες φυσικές υποθέσεις.

Τα τελευταία 30 χρόνια, η ανάπτυξη της Στατιστικής Μάθησης έχει συνδεθεί με τη μελέτη ενιαίων νόμων μεγάλων αριθμών. Η θεωρία παρείχε μια κατανόηση των εγγενών περιπλοκών της μάθησης χωρίς διανομές, καθώς και των περιορισμένων δειγμάτων και των εξαρτώμενων από τα δεδομένα εγγυήσεων. Εκτός από την καθιερωμένη θεωρία, οι αλγόριθμοι

που αναπτύσσονται από την κοινότητα θεωρούνται συχνά ως οι πλέον σύγχρονες μέθοδοι για προβλήματα πρόβλεψης.

Αυτές οι μέθοδοι τηρούν τη φιλοσοφία, για παράδειγμα, τα προβλήματα ταξινόμησης δεν πρέπει να μοντελοποιούν τις διανομές αλλά μάλλον να διαμορφώνουν το όριο της απόφασης. Αναμφισβήτητα, αυτό αντιπροσωπεύει την επιτυχία πολλών μεθόδων μάθησης, με το μειονέκτημα ότι η ερμηνεία των αποτελεσμάτων είναι συχνά πιο δύσκολη. Ο όρος «μάθηση» είναι μια κληρονομιά της ισχυρής σύνδεσης του τομέα με τα προβλήματα που οφείλονται στον υπολογιστή και επισημαίνει ότι ο στόχος δεν είναι απαραιτήτως ο «υπολογισμός της πραγματικής παραμέτρου», αλλά μάλλον η βελτίωση της απόδοσης με περισσότερα δεδομένα[22].

Σε γενικές γραμμές, η εξόρυξη δεδομένων περιλαμβάνει τεχνικές και αλγόριθμους για τον προσδιορισμό ενδιαφερόντων σχεδίων από μεγάλα σύνολα δεδομένων. Υπάρχουν επί του παρόντος εκατοντάδες αλγόριθμοι που εκτελούν καθήκοντα όπως η συχνή εξόρυξη προτύπων, η ομαδοποίηση και η ταξινόμηση, μεταξύ άλλων. Η κατανόηση του τρόπου με τον οποίο αυτοί οι αλγόριθμοι λειτουργούν και του πώς μπορούν να χρησιμοποιηθούν αποτελεσματικά είναι μια συνεχής πρόκληση που αντιμετωπίζουν οι αναλυτές εξόρυξης δεδομένων, οι ερευνητές και οι επαγγελματίες, ιδίως επειδή η συμπεριφορά και τα πρότυπα αλγορίθμων που παρέχει μπορεί να αλλάξουν σημαντικά ως συνάρτηση των παραμέτρων της. Υπάρχει ένας μεγάλος αριθμός διαθέσιμων υλοποιήσεων(όπως εκείνοι στο R).

## 1.3 Εισαγωγή στην Ανάλυση Δεδομένων

Τα δεδομένα χρησιμοποιούνται για να περιγράψουν τα πράγματα με την ανάθεση μιας αξίας σε αυτά. Οι αξίες οργανώνονται, επεξεργάζονται και παρουσιάζονται μέσα σε ένα δεδομένο πλαίσιο έτσι ώστε να καταστούν χρήσιμες.

**Τα δεδομένα μπορούν να έχουν διάφορες μορφές:**

Ποιοτικά δεδομένα

Τα "ποιοτικά δεδομένα" είναι δεδομένα που χρησιμοποιούν λέξεις και περιγραφές. Μπορούν να παρατηρηθούν ποιοτικά δεδομένα, αλλά είναι υποκειμενικά και ως εκ τούτου δύσκολο να χρησιμοποιηθούν για σκοπούς σύγκρισης. Περιγραφές υφής, γεύσης ή εμπειρίας αποτελούν παραδείγματα ποιοτικών δεδομένων. Οι ποιοτικές μέθοδοι συλλογής δεδομένων περιλαμβάνουν ομάδες εστίασης, συνεντεύξεις ή αντικείμενα ανοιχτού τύπου σε μια έρευνα.

Ποσοτικά δεδομένα

Τα "ποσοτικά δεδομένα" είναι δεδομένα που εκφράζονται με αριθμούς. Τα ποσοτικά δεδομένα είναι δεδομένα που μπορούν να ταξινομηθούν, να μετρηθούν ή να ταξινομηθούν. Το μήκος, το βάρος, η ηλικία, το κόστος, οι κλίμακες αξιολόγησης, είναι όλα παραδείγματα ποσοτικών δεδομένων. Τα ποσοτικά δεδομένα μπορούν να αναπαρασταθούν οπτικά σε γραφήματα και πίνακες και να αναλυθούν στατιστικά.

**Υπάρχουν δύο είδη ποσοτικών δεδομένων:**

Κατηγορικά δεδομένα

Τα "κατηγορικά δεδομένα" είναι δεδομένα που έχουν τοποθετηθεί σε ομάδες. Ένα στοιχείο δεν μπορεί να ανήκει σε περισσότερες από μία ομάδες τη φορά. Παραδείγματα κατηγορικών δεδομένων είναι η τρέχουσα κατάσταση διαβίωσης του ατόμου, το καθεστώς καπνίσματος ή εάν απασχολείται.

Συνεχή δεδομένα

Ως "συνεχή δεδομένα" νοούνται τα αριθμητικά δεδομένα με συνεχή κλίμακα. Στα συνεχή δεδομένα, όλες οι τιμές είναι δυνατές χωρίς κενά μεταξύ τους. Παραδείγματα συνεχών δεδομένων είναι το ύψος ή το βάρος του ατόμου και η θερμοκρασία.

Η **Ανάλυση Δεδομένων** (*data analysis*), αποτελεί στην ουσία μια διαδικασία συλλογής (παρατήρηση, απόκτηση), επεξεργασίας (καθαρισμός, μετατροπή) και μοντελοποίησης των

δεδομένων με στόχο την εξεύρεση χρήσιμης πληροφορίας για την υποστήριξη διαφόρων ειδών λήψεων αποφάσεων (decision-making). Η ¨εξόρυξη γνώσης ¨ είναι μια συγκεκριμένη υποκατηγορία ή αλλιώς τεχνική ανάλυσης που επικεντρώνεται στην εύρεση μοντέλων (προτύπων) και την αναζήτηση γνώσης σκοπεύοντας κυρίως στην πρόβλεψη και όχι στην περιγραφή φαινομένων και συμπεριφορών.

Η προγνωστική ανάλυση (predictive analytics), στοχεύει στην εφαρμογή στατιστικών μοντέλων για την πρόβλεψη ή κατηγοριοποίηση δεδομένων, καθώς επίσης και η ανάλυση κειμένου (text analytics) επιτυγχάνεται με την εφαρμογή στατιστικών εργαλείων σε συνδυασμό με γλωσσολογικές τεχνικές ώστε να εξαχθεί και να κατηγοριοποιηθεί πληροφορία από πήγες με δεδομένα χωρίς δομή (unstructured data).

Ο μεγάλος στατιστικός John Tukey [23] όρισε για πρώτη φορά την ευρύτερη επιστήμη της Ανάλυσης Δεδομένων ως τις:

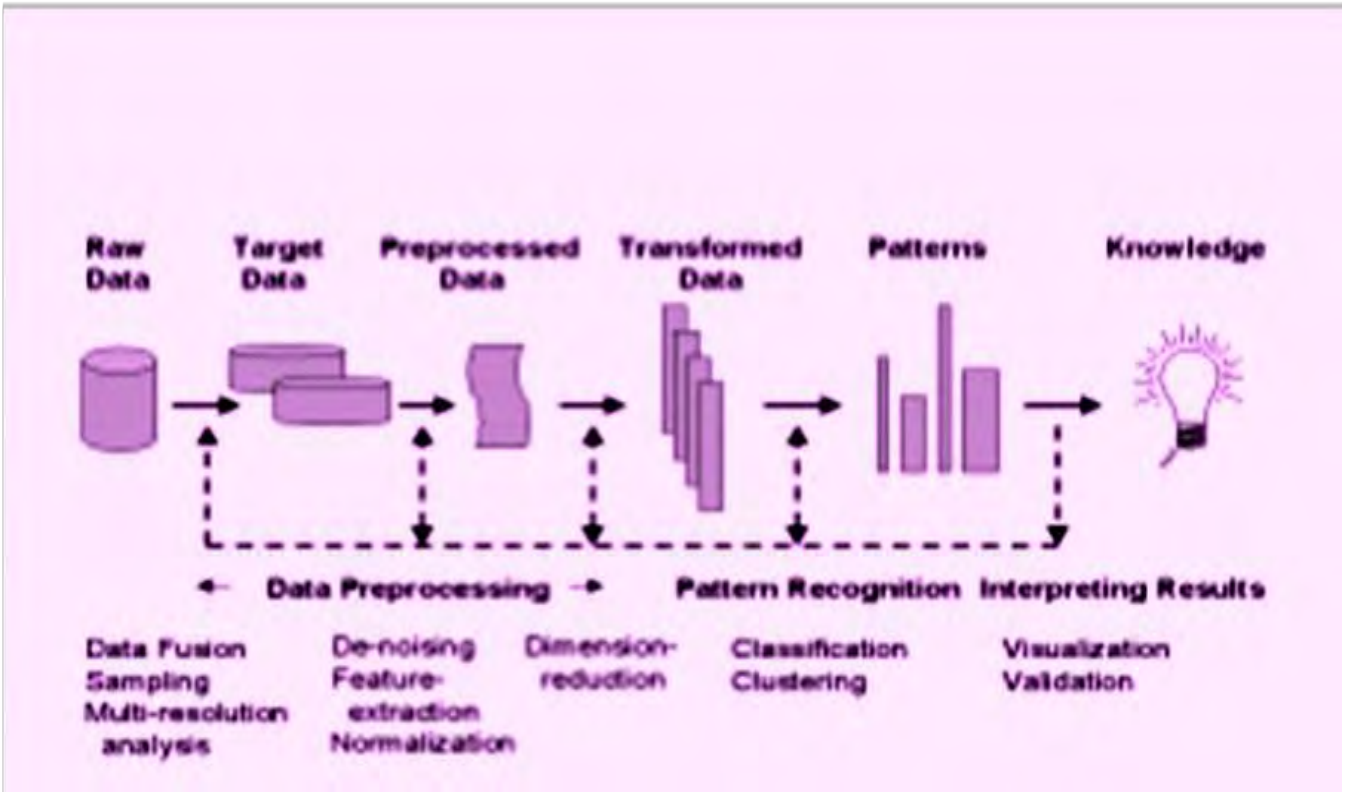
*“Διαδικασίες για την ανάλυση δεδομένων, τεχνικές για την διερμηνεία των αποτελεσμάτων τέτοιων διαδικασιών, τρόποι οργάνωσης για την συλλογή δεδομένων ώστε να κάνουν την ανάλυση ευκολότερη, πιο συγκεκριμένη και με μεγαλύτερη ακρίβεια, και όλη η μηχανική σε συνδυασμό με τα αποτελέσματα μαθηματικών συναρτήσεων και στατιστικών μέθοδών τα οποία εφαρμόζονται για την ανάλυση των δεδομένων.”*

## 2. Εξόρυξη Δεδομένων (Data Mining)

Εξόρυξη δεδομένων (ή ανακάλυψη γνώσης από βάσεις δεδομένων) είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Ο όρος εξόρυξη δεδομένων είναι μία έννοια που συνήθως παραπέμπει σε κάθε είδος φόρμας με μεγάλη ποσότητα δεδομένων ή επεξεργασία δεδομένων (συλλογή, εξαγωγή δεδομένων, warehouse, ανάλυση δεδομένων και στατιστικής) αλλά επίσης γενικεύεται σε κάθε είδος συστήματος υποστήριξης αποφάσεων συμπεριλαμβανομένου της τεχνητής νοημοσύνης, της εκμάθησης μηχανής και της επιχειρηματικής ευφυΐας. Στην ορθή χρήση του όρου η λέξη κλειδί είναι η ανακάλυψη, που ορίζεται ως η ανίχνευση κάτι καινούριου.

Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

Η εξόρυξη δεδομένων είναι μια διαδικασία που αφορά την αποκάλυψη προτύπων, συσχετίσεων, ανωμαλιών και στατιστικά σημαντικών δομών στα δεδομένα[24]. Συνήθως αναφέρεται στην περίπτωση όπου τα δεδομένα είναι πολύ μεγάλα ή πολύ περίπλοκα για να επιτρέπουν είτε χειροκίνητη ανάλυση είτε ανάλυση μέσω απλών ερωτημάτων. Η εξόρυξη δεδομένων αποτελείται από την προ επεξεργασία δεδομένων, κατά την οποία εξάγονται από τα δεδομένα χαμηλού επιπέδου σχετικές υψηλού επιπέδου χαρακτηριστικά ή ιδιότητες και αναγνώριση προτύπου, στο οποίο αναγνωρίζεται ένα πρότυπο στα δεδομένα χρησιμοποιώντας αυτά τα χαρακτηριστικά. (Σχήμα 1). Η προ επεξεργασία των δεδομένων είναι συχνά ένα χρονοβόρο, αλλά κρίσιμο, πρώτο βήμα. Για να εξασφαλιστεί η επιτυχία της διαδικασίας εξόρυξης δεδομένων, είναι σημαντικό τα χαρακτηριστικά που εξάγονται από τα δεδομένα να είναι σχετικά με το πρόβλημα και αντιπροσωπευτικά των δεδομένων.




Σχήμα 1: Data Mining – an iterative and interactive process

Ανάλογα με τον τύπο των δεδομένων που εξ ορύσσονται, το βήμα προ επεξεργασίας μπορεί να αποτελείται από πολλές υπό εργασίες. Εάν τα ανεπεξέργαστα δεδομένα είναι πολύ μεγάλα, θα μπορούσαμε να χρησιμοποιήσουμε δειγματοληψία και να δουλέψουμε με λιγότερες περιπτώσεις, ή να χρησιμοποιήσουμε τεχνικές πολλαπλής ανάλυσης και να δουλέψουμε με δεδομένα με μεγαλύτερη ανάλυση. Στη συνέχεια, ο θόρυβος των δεδομένων αφαιρείται στο μέτρο του δυνατού και εξάγονται τα σχετικά χαρακτηριστικά. Σε ορισμένες περιπτώσεις, όπου είναι διαθέσιμα δεδομένα από διαφορετικές πηγές ή αισθητήρες, μπορεί να απαιτηθεί σύντηξη δεδομένων για να επιτραπεί η εκμετάλλευση σε όλα τα διαθέσιμα δεδομένα για ένα πρόβλημα. Στο τέλος αυτού του πρώτου βήματος, έχουμε ένα διάνυσμα χαρακτηριστικών για κάθε περίπτωση δεδομένων.

Ανάλογα με το πρόβλημα και τα δεδομένα, ίσως χρειαστεί να μειώσουμε τον αριθμό των χαρακτηριστικών χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών ή τεχνικές μείωσης των διαστάσεων ή οι μη γραμμικές εκδοχές του[25]. Μετά από αυτήν την προ επεξεργασία, τα δεδομένα είναι έτοιμα για την ανίχνευση μοτίβων μέσω της χρήσης αλγορίθμων όπως

ταξινόμηση, ομαδοποίηση, παλινδρόμηση κλπ. Αυτά τα μοτίβα εμφανίζονται στη συνέχεια στον χρήστη για επικύρωση. Η εξόρυξη δεδομένων είναι μια επαναληπτική και δια δραστική διαδικασία.

Η εξόρυξη δεδομένων είναι ένα πρακτικό θέμα και περιλαμβάνει τη μάθηση με πρακτική και όχι θεωρητική έννοια. Ενδιαφερόμαστε για τεχνικές για την εύρεση και περιγραφή δομικών προτύπων στα δεδομένα ως εργαλείο για να βοηθήσουμε να εξηγήσουμε αυτά τα δεδομένα και να κάνουμε προβλέψεις από αυτό. Τα δεδομένα θα λάβουν τη μορφή μιας σειράς παραδειγμάτων. Η έξοδος παίρνει τη μορφή προβλέψεων για νέα παραδείγματα. Η έξοδος μπορεί επίσης να περιλαμβάνει μια πραγματική περιγραφή μιας δομής που μπορεί να χρησιμοποιηθεί για να ταξινομήσει άγνωστα παραδείγματα για να εξηγήσει την απόφαση. Εκτός από τις επιδόσεις, είναι χρήσιμο να παρέχεται μια ρητή αναπαράσταση της γνώσης που αποκτάται. Πολλές τεχνικές μάθησης αναζητούν διαρθρωτικές περιγραφές του τι μαθαίνεται, περιγραφές που μπορούν να γίνουν αρκετά περίπλοκες και τυπικά εκφράζονται ως σύνολα κανόνων ή τα δέντρα απόφασης που περιγράφονται αργότερα. Επειδή μπορούν να γίνουν κατανοητά από τους ανθρώπους, οι περιγραφές αυτές χρησιμεύουν για να εξηγήσουν τι έχει μάθει και να εξηγήσουν τη βάση για νέες προλήψεις. Η εμπειρία δείχνει ότι σε πολλές εφαρμογές μηχανικής μάθησης στην εξόρυξη δεδομένων, οι δομές ρητής γνώσης που αποκτώνται, οι δομικές περιγραφές είναι τουλάχιστον εξίσου σημαντικές και συχνά πολύ πιο σημαντικές από την ικανότητα να αποδίδουν καλά σε νέα παραδείγματα. Οι άνθρωποι συχνά χρησιμοποιούν εξόρυξη δεδομένων για να αποκτήσουν γνώση, όχι μόνο προβλέψεις.

Η χρήση δεδομένων - ιδίως δεδομένων για τους ανθρώπους - για την εξόρυξη δεδομένων έχει σοβαρές δεοντολογικές συνέπειες και οι επαγγελματίες των τεχνικών εξόρυξης δεδομένων πρέπει να ενεργούν υπεύθυνα ενημερώνοντάς τους για τα δεοντολογικά ζητήματα που περιβάλλουν τη συγκεκριμένη εφαρμογή τους.

Είναι ευρέως αποδεκτό ότι, πριν οι άνθρωποι λάβουν απόφαση να παράσχουν προσωπικές πληροφορίες πρέπει να γνωρίζουν πώς θα χρησιμοποιηθούν και σε τι θα χρησιμοποιηθούν, τι μέτρα θα ληφθούν για την προστασία της εμπιστευτικότητας και της ακεραιότητάς τους, ποιες είναι οι συνέπειες της παροχής ή παρακράτηση των πληροφοριών, καθώς και τυχόν δικαιώματά τους. Κάθε φορά που συλλέγονται τέτοιες πληροφορίες, τα άτομα πρέπει να τους λένε αυτά τα πράγματα.

Η πιθανή χρήση τεχνικών εξόρυξης δεδομένων σημαίνει ότι οι τρόποι με τους οποίους μπορεί να χρησιμοποιηθεί ένα αποθετήριο δεδομένων μπορεί να εκτείνεται πολύ πέρα από αυτό που σχεδιάστηκε όταν αρχικά συλλέχθηκαν τα δεδομένα. Αυτό δημιουργεί ένα σοβαρό πρόβλημα. Είναι απαραίτητο να καθοριστούν οι συνθήκες υπό τις οποίες συλλέχθηκαν τα δεδομένα και για ποιους σκοπούς μπορούν να χρησιμοποιηθούν. Σαφώς στην περίπτωση των προσωπικών δεδομένων που συλλέγονται ρητώς, η ιδιοκτησία των δεδομένων δεν απονέμει το δικαίωμα χρήσης των δεδομένων με διαφορετικούς τρόπους από αυτούς που προβάλλονταν κατά την αρχική καταγραφή τους.

Εκπληκτικά πράγματα εξέρχονται από την εξόρυξη δεδομένων[26].

Χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων για την ανάλυση δεδομένων σε διάφορους τομείς, όπως η τηλεπισκόπηση, η βιοπληροφορική, η ιατρική απεικόνιση, η αστρονομία, η εξόρυξη ιστού, η εξόρυξη κειμένου, η διαχείριση σχέσεων με τους πελάτες και η ανάλυση του καλαθιού αγοράς. Ενώ ένα μεγάλο μέρος της εστίασης στη διαδικασία εξόρυξης δεδομένων τείνει να είναι σε αλγόριθμους αναγνώρισης προτύπων, τα βήματα προ επεξεργασίας των δεδομένων έχουν μεγαλύτερη επιρροή στην επιτυχία της προσπάθειας εξόρυξης δεδομένων[27]. Τα βήματα προ επεξεργασίας συχνά εξαρτώνται από τον τομέα και το πρόβλημα.

Η εξόρυξη δεδομένων βασίζεται στην αυτόματη ή ημι-αυτόματη ανακάλυψη μοντέλων και βασίζεται σε αλγόριθμους εκπαίδευσης Η/Υ, όπως τα νευρωνικά δίκτυα (για μη-γραμμικούς συσχετισμούς) και οι γενετικοί αλγόριθμοι (που προσομοιάζουν την διαδικασία της φυσικής εξέλιξης).

Ενώ η κλασσική στατιστική προϋποθέτει ότι το σύνολο των προς επεξεργασία δεδομένων θα βρίσκονται στη μνήμη του Η/Υ, αυτό συνήθως δεν είναι εφικτό και απαιτούνται τεχνικές συσχετισμού που διαχειρίζονται μεγάλες βάσεις δεδομένων.

## 2.1 Κατηγοριοποίηση μεθόδων εξόρυξης δεδομένων:

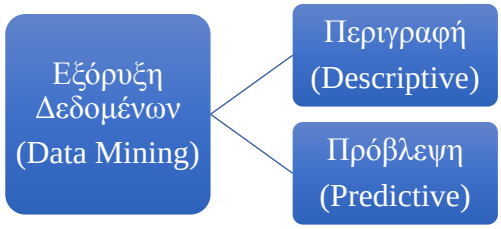
Είναι πολύ σημαντικό να οριστεί η εργασία εξόρυξης δεδομένων που πρέπει να αντιμετωπίσει ο αλγόριθμος προτού σχεδιαστεί για εφαρμογή σε ένα συγκεκριμένο πρόβλημα. Υπάρχουν διάφορα καθήκοντα εξόρυξης δεδομένων και κάθε ένας από αυτούς έχει συγκεκριμένους σκοπούς όσον αφορά τις γνώσεις που πρέπει να ανακαλυφθούν.

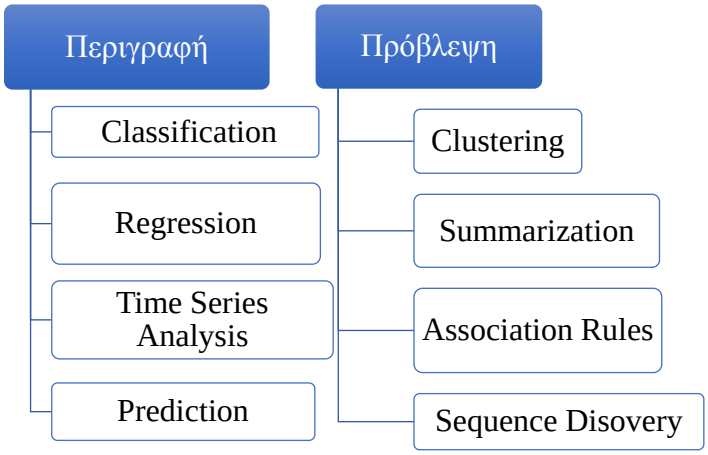
Στην εξόρυξη δεδομένων ο όρος μοντέλο "είναι μια περιγραφή υψηλού επιπέδου του συνόλου δεδομένων". Ένα μοντέλο μπορεί να είναι περιγραφικό ή προγνωστικό. Όπως υποδηλώνουν τα ονόματα, το περιγραφικό μοντέλο είναι ένα μη εποπτευόμενο μοντέλο που στοχεύει να περιγράψει τα δεδομένα, ενώ το μοντέλο πρόβλεψης είναι ένα εποπτευόμενο μοντέλο που στοχεύει στην πρόβλεψη αξιών από τα δεδομένα.

Τα μοτίβα χρησιμοποιούνται για τον καθορισμό των σημαντικών και ενδιαφερόντων χαρακτηριστικών των δεδομένων. Τα μοντέλα χρησιμοποιούνται για την περιγραφή ολόκληρου του συνόλου δεδομένων, ενώ τα μοτίβα χρησιμοποιούνται για την επισήμανση συγκεκριμένων πτυχών των δεδομένων.

Η ανάλυση δεδομένων μπορεί γενικά να είναι είτε σε μορφή ταξινόμησης είτε σε μορφή πρόβλεψης. Η ανάλυση παλινδρόμησης είναι ένα παράδειγμα των καθηκόντων πρόβλεψης, δηλαδή της αριθμητικής πρόβλεψης. Η διαφορά μεταξύ της ταξινόμησης και της παλινδρόμησης είναι ότι η τιμή-στόχος (μεταβλητή απόκρισης) είναι μια ποσοτική τιμή στη μοντελοποίηση παλινδρόμησης, ενώ στην ταξινόμηση μοντέλων είναι μια ποιοτική ή κατηγορική τιμή[28].

Υπάρχει μια μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων. Ανάλογα με το είδος των δεδομένων και το είδος της γνώσης που εξάγεται, αυτές κατηγοριοποιούνται σε διαφορετική́ κατηγορία. Μερικές βασικές μέθοδοι της Εξόρυξης Δεδομένων παρουσιάζονται παρακάτω.




**Κατηγοριοποίηση - Ταξινόμηση (Classification)**: Πρόκειται για μία προγνωστική μέθοδο. Στόχος είναι η δημιουργία ενός μοντέλου-κατηγοριοποιητή (classifier) με βάση τα υπάρχοντα δεδομένα. Ουσιαστικά, είναι η μάθηση μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο (συνήθως αναπαρίσταται ως ένα διάνυσμα τιμών για τις χαρακτηριστικές του ιδιότητες) σε μία τιμή μιας κατηγορικής μεταβλητής, η οποία είναι γνωστή και ως κλάση(ή κατηγορία).

Η κατηγοριοποίηση συχνά συγχέεται με το γενικό όρο της πρόβλεψης. Στην κατηγοριοποίηση, το αποτέλεσμα που θέλουμε να προβλέψουμε είναι η κλάση των δειγμάτων. Η κλάση μπορεί να πάρει διακριτές τιμές από ένα πεπερασμένο σύνολο. Αντίθετα, κατά την πρόβλεψη με χρήση τεχνικών όπως η παλινδρόμηση, η μεταβλητή-στόχος μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός.

**Παλινδρόμηση (Regression):** Μια σχετική διαδικασία με την κατηγοριοποίηση είναι η παλινδρόμηση, στόχος της οποίας είναι η μάθηση ή αλλιώς η εκπαίδευση (*training*) μιας

συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή. Πρόκειται για μια, επίσης, προγνωστική μέθοδο. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable). Χρησιμοποιείται σαν τεχνική εδώ και αιώνες με πιο γνωστές μεθόδους την γραμμική (linear) και την λογιστική (logistic) παλινδρόμηση. Άλλες τεχνικές περιλαμβάνουν τα δένδρα παλινδρόμησης (regression trees) και τα νευρωνικά δίκτυα (neural networks). Εάν η μεταβλητή στόχευσης δεν είναι συνεχής αριθμός, τότε εφαρμόζονται τεχνικές όπως η λογιστική παλινδρόμηση, για την οποία θα γίνει αναφορά στη συνέχεια.

**Συσταδοποίηση (Clustering):** Η συσταδοποίηση είναι μια περιγραφική μέθοδος. Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων, δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα. Ουσιαστικά αναζητείται ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, για να περιγράψει τα δεδομένα. Οι κατηγορίες μπορεί να είναι αμοιβαία αποκλεισμένες και εξαντλητικές ή να έχουν μία πιο σύνθετη αναπαράσταση, όπως για παράδειγμα ιεραρχικές και επικαλυπτόμενες.

**Εξαγωγή και Ανάλυση Συσχετίσεων (Association Rules)**: Η εξαγωγή κανόνων συσχέτισης θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Έχει προσελκύσει ιδιαίτερο ενδιαφέρον, καθώς οι κανόνες συσχέτισης παρέχουν έναν συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται ευκολά κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στη μορφή Α → Β, όπου τα Α και Β αποτελούν σύνολα που αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων που θέλουμε να αναλύουμε. Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης Α→Β προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου Β δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου Α.

***Οδηγίες για μία επιτυχημένη εξόρυξη δεδομένων:***

- Τα δεδομένα πρέπει να είναι διαθέσιμα.

- Τα δεδομένα πρέπει να είναι σχετικά με τον στόχο της εξόρυξης δεδομένων και αρκετά για την τεχνική που πρέπει να εφαρμόσουμε, να έχει εφαρμοστεί μια διαδικασία «καθαρισμού».
- Το πρόβλημα θα πρέπει να είναι καθορισμένο.
- Το πρόβλημα δεν θα πρέπει να επιλύεται με τη χρήση ερωτημάτων SQL ή OLAP.
- Τα αποτελέσματα θα πρέπει να οδηγούν σε ενέργειες.

## 3. Μηχανική Μάθηση (Machine Learning)

Ο Arthur Samuel είναι ένας από τους πρωτοπόρους της μηχανικής μάθησης. Χαρακτήρισε την εκμάθηση μηχανών ως "*πεδίο σπουδών που δίνει τη δυνατότητα στους υπολογιστές να μάθουν χωρίς να έχουν προγραμματιστεί ρητά*".

Η μηχανική μάθηση είναι ένας τομέας της επιστήμης των υπολογιστών. Είναι επίσης ένας τύπος Τεχνητής Νοημοσύνης που επιτρέπει στους προγραμματιστές να γράφουν προγράμματα με πιο απλό τρόπο. Επικεντρώνεται περισσότερο στην ανάπτυξη προγραμμάτων που διδάσκουν τους υπολογιστές να αλλάζουν όταν εκτίθενται σε νέα δεδομένα και αναπτύσσονται. Σκοπός του είναι να κατανοήσει και να ακολουθήσει τις μεθόδους χρησιμοποιώντας αλγορίθμους για αυτόματη εκτέλεση του έργου χωρίς ανθρώπινη βοήθεια[29].

Ο Tom Mitchell, ένας ακόμη γνωστός ερευνητής μηχανικής μάθησης, πρότεινε έναν πιο ακριβή ορισμό το 1998: *Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σε σχέση με κάποιο έργο T και κάποιο μέτρο απόδοσης P, αν η απόδοσή του στο T, όπως μετράται από το P, βελτιώνεται με την εμπειρία E.*

Η μηχανική μάθηση βασίζεται σε ιδέες από μια ποικιλία επιστημονικών κλάδων, όπως η τεχνητή νοημοσύνη, η πιθανότητα και οι στατιστικές, η υπολογιστική πολυπλοκότητα, η θεωρία της πληροφορίας, η ψυχολογία και η νευροβιολογία, η θεωρία ελέγχου και η φιλοσοφία.
Ένα καλά καθορισμένο μαθησιακό πρόβλημα απαιτεί μια καλά καθορισμένη εργασία, μετρική απόδοση και πηγή εμπειρίας κατάρτισης.

Ο σχεδιασμός μιας προσέγγισης μηχανικής μάθησης περιλαμβάνει διάφορες επιλογές σχεδιασμού, συμπεριλαμβανομένης της επιλογής του τύπου της εκπαιδευτικής εμπειρίας, της λειτουργίας στόχου που πρέπει να μάθει, μιας αναπαράστασης αυτής της λειτουργίας στόχου και ενός αλγορίθμου για την εκμάθηση της λειτουργίας στόχου από παραδείγματα εκπαίδευσης.
Η εκμάθηση περιλαμβάνει αναζήτηση μέσα από ένα χώρο πιθανών υποθέσεων για να βρεθεί η υπόθεση που ταιριάζει καλύτερα στα διαθέσιμα παραδείγματα εκπαίδευσης και σε άλλους προηγούμενους περιορισμούς ή γνώσεις[30].

Η μηχανική μάθηση διερευνά τον τρόπο με τον οποίο οι υπολογιστές μπορούν να μάθουν (ή να βελτιώσουν την απόδοσή τους) βάσει δεδομένων. Ένας κύριος τομέας έρευνας είναι τα προγράμματα ηλεκτρονικών υπολογιστών να μάθουν αυτόματα να αναγνωρίζουν πολύπλοκα πρότυπα και να λαμβάνουν έξυπνες αποφάσεις βασισμένες σε δεδομένα.
Η μηχανική μάθηση είναι μια ταχέως αναπτυσσόμενη πειθαρχία[31].

Οι αλγόριθμοι μηχανικής μάθησης έχουν αποδειχθεί ότι έχουν μεγάλη πρακτική αξία σε μια ποικιλία τομέων εφαρμογής.

Είναι ιδιαίτερα χρήσιμοι σε:

- Προβλήματα εξόρυξης δεδομένων όπου μεγάλες βάσεις δεδομένων μπορεί να περιέχουν πολύτιμες κανονικότητες που μπορούν να ανακαλυφθούν αυτόματα (π.χ., να αναλύσουν τα αποτελέσματα των ιατρικών θεραπειών από τις βάσεις δεδομένων των ασθενών ή να μάθουν γενικούς κανόνες για την αξιοπιστία από τις οικονομικές βάσεις δεδομένων).

- Ανεπαρκώς κατανοητούς τομείς όπου οι άνθρωποι μπορεί να μην έχουν τις γνώσεις που απαιτούνται για την ανάπτυξη αποτελεσματικών αλγορίθμων (π.χ. αναγνώριση ανθρώπινου προσώπου από εικόνες).
- Τομείς στους οποίους το πρόγραμμα πρέπει να προσαρμόζεται δυναμικά στις μεταβαλλόμενες συνθήκες (π.χ. έλεγχος των παραγωγικών διαδικασιών στο πλαίσιο μεταβαλλόμενων αποθεμάτων εφοδιασμού ή προσαρμογή στα μεταβαλλόμενα ενδιαφέροντα ανάγνωσης των ατόμων.

Από την μελέτη της αναγνώρισης προτύπων και της θεωρίας της υπολογιστικής μάθησης στην τεχνητή νοημοσύνη, η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μάθουν και να κάνουν προβλέψεις στα δεδομένα. Οι αλγόριθμοι αυτοί λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση χρησιμοποιείται σε μια σειρά υπολογιστικών εργασιών, όπου ο σχεδιασμός και ο προγραμματισμός σαφών αλγορίθμων με καλές επιδόσεις είναι δύσκολος ή μη εφικτός. Συνδέεται στενά με (και συχνά επικαλύπτει) υπολογιστικές στατιστικές και επικεντρώνεται επίσης στην πρόβλεψη μέσω της χρήσης υπολογιστών. Έχει ισχυρούς δεσμούς με τη μαθηματική βελτιστοποίηση, η οποία παρέχει τις μεθόδους, τη θεωρία και τους τομείς εφαρμογής.

Η μηχανική μάθηση συνδυάζεται με την εξόρυξη δεδομένων (data mining), όπου αυτός ο τομέας εστιάζει περισσότερο στην ανάλυση διερευνητικών δεδομένων και είναι γνωστός ως μάθηση χωρίς επίβλεψη. Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την ανάπτυξη σύνθετων μοντέλων και αλγορίθμων που προσφέρονται για την πρόβλεψη. Αυτά τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες, τους μηχανικούς και τους αναλυτές να ¨παράγουν αξιόπιστες και επαναλαμβανόμενες αποφάσεις και αποτελέσματα¨ και να αποκαλύπτουν ¨κρυφές ιδέες¨ μέσω της μάθησης από τις ιστορικές σχέσεις και τις τάσεις στα δεδομένα.

Με την άνοδο των μεγάλων δεδομένων, η μηχανική μάθηση έχει καταστεί βασική τεχνική για την επίλυση προβλημάτων σε τομείς όπως :

- Υπολογιστική χρηματοδότηση, για πιστωτική βαθμολογία και αλγοριθμική διαπραγμάτευση.
- Επεξεργασία εικόνας και όραση υπολογιστή, για αναγνώριση προσώπου, ανίχνευση κίνησης και ανίχνευση αντικειμένων.
- Υπολογιστική βιολογία, για ανίχνευση όγκων, ανακάλυψη φαρμάκου και αλληλουχία DNA.
- Παραγωγή ενέργειας, για την πρόβλεψη των τιμών και του φορτίου.
- Αυτοκινητοβιομηχανία, αεροδιαστημική και κατασκευή, για προληπτική συντήρηση.
- Επεξεργασία φυσικής γλώσσας, για εφαρμογές φωνητικής αναγνώρισης.

**Πλεονεκτήματα της Μηχανικής Μάθησης:**

Η μηχανική μάθηση χρησιμοποιείται όλο και περισσότερο και υπάρχουν πολλοί λόγοι για αυτό.

1. Το μηχάνημα βελτιώνεται από τα δικά του λάθη.
2. Ταχύτερη από τους ανθρώπους, έτσι εξοικονομείται χρόνος.
3. Χρειάστηκαν αρκετά χρόνια εργασίας πάνω σε αυτό, για να είναι σωστό από πολλές παραμέτρους.
4. Χρησιμοποιείται ευρέως στην αναγνώριση προσώπου και επίσης σε φωτογραφίες με υπότιτλους. Χρησιμοποιείται για να συλλάβει κλέφτες ή όποιον εμπλέκεται σε μια σκηνή εγκλήματος από την κάμερα.

**Μειονεκτήματα της Μηχανικής Μάθησης:**

Δεν είναι εγγυημένο ότι οι αλγόριθμοι μηχανικής μάθησης θα λειτουργούν πάντα για κάθε περίπτωση. Μερικές φορές η μηχανική μάθηση θα αποτύχει, οπότε πρέπει να καταλάβουμε το πρόβλημα για να εφαρμοστεί ο αλγόριθμος μηχανικής μάθησης με τον σωστό τρόπο. Ορισμένοι αλγόριθμοι μηχανικής μάθησης απαιτούν πολλά δεδομένα, όπως αλγόριθμοι βαθιάς μάθησης. Μπορεί να είναι δύσκολο να κερδίσετε αυτό το μεγάλο όγκο δεδομένων, όμως υπάρχουν πολλά δεδομένα για ιατρικούς, εκπαιδευτικούς σκοπούς και αναγνώριση εικόνας.

3.1 Είδη Μηχανικής Μάθησης*:*

- Η επιτηρούμενη μάθηση
- Η μη επιτηρούμενη μάθηση
- Η ενισχυτική μάθηση

Η **επιτηρούμενη μάθηση**, χτίζει ένα μοντέλο που κάνει τις προβλέψεις βασισμένες σε αποδεικτικά στοιχεία παρουσία αβεβαιότητας. Ένας επιτηρούμενος αλγόριθμος εκμάθησης παίρνει ένα γνωστό σύνολο δεδομένων εισόδου και γνωστών απαντήσεων στα δεδομένα (έξοδος) και εκπαιδεύει ένα μοντέλο για να παράγει λογικές προβλέψεις για την απάντηση σε νέα δεδομένα. Η χρήση επιτηρούμενης μάθησης γίνεται εάν γνωρίζουμε δεδομένα για την παραγωγή που προσπαθούμε να προβλέψουμε. Χρησιμοποιείται σε προβλήματα ταξινόμησης, πρόγνωσης και διερμηνείας.
Χρησιμοποιεί τεχνικές ταξινόμησης και παλινδρόμησης για την ανάπτυξη προγνωστικών μοντέλων.

Οι τεχνικές ταξινόμησης (classification techniques), προβλέπουν διακριτές απαντήσεις. Για παράδειγμα, εάν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι γνήσιο ή ανεπιθύμητο, ή αν ένας όγκος είναι καρκίνος ή καλοήθεις. Τα μοντέλα ταξινόμησης ταξινομούν τα δεδομένα εισόδου σε κατηγορίες. Τυπικές εφαρμογές περιλαμβάνουν την ιατρική απεικόνιση, την αναγνώριση ομιλίας και τη βαθμολόγηση της πίστωσης. Χρήση ταξινόμησης γίνεται εάν τα δεδομένα μπορούν να επισημανθούν, να κατηγοριοποιηθούν ή να διαχωριστούν σε συγκεκριμένες ομάδες ή κλάσεις. Για παράδειγμα, οι εφαρμογές αναγνώρισης χεριών χρησιμοποιούν ταξινόμηση για να αναγνωρίζουν γράμματα και αριθμούς. Στην επεξεργασία εικόνας και την όραση του υπολογιστή, οι τεχνικές αναγνώρισης μοτίβων χωρίς εποπτεία χρησιμοποιούνται για ανίχνευση αντικειμένων και τμηματοποίηση εικόνας. Οι συνήθεις αλγόριθμοι για την εκτέλεση της ταξινόμησης περιλαμβάνουν τη μηχανή διανύσματος υποστήριξης (support vector machine, SVM), τα ενισχυμένα δέντρα αποφάσεων (boosted and bagged decision trees), τον πλησιέστερο γείτονα (k-nearest neighbor), τον Naive Bayes, την διακριτική ανάλυση (distriminant analysis), την λογιστική παλινδρόμηση (logistic regression) και τα νευρωνικά δίκτυα (neural networks).

Οι τεχνικές παλινδρόμησης (regression techniques), προβλέπουν συνεχείς αποκρίσεις. Για παράδειγμα, μεταβολές στη θερμοκρασία ή διακυμάνσεις της ζήτησης ισχύος. Χρήση τεχνικών παλινδρόμησης γίνεται εάν εργάζεστε με ένα εύρος δεδομένων ή εάν η φύση της απάντησής σας είναι ένας πραγματικός αριθμός. Οι συνήθεις αλγόριθμοι παλινδρόμησης περιλαμβάνουν το γραμμικό μοντέλο (linear model), το μη γραμμικό μοντέλο (nonlinear model), την τακτοποίηση (regularization), τη σταδιακή παλινδρόμηση ( stepwise regression), τα ενισχυμένα δέντρα αποφάσεων (boosted and bagged decision trees), τα νευρωνικά δίκτυα ( neural networks) και την προσαρμοζόμενη νεύρο-ασαφή μάθηση (adaptive neuro-fuzzy learning).

Η **μη επιτηρούμενη μάθηση**, βρίσκει κρυμμένα μοτίβα ή εγγενείς δομές στα δεδομένα. Χρησιμοποιείται για την εξαγωγή συμπερασμάτων από σύνολα δεδομένων που αποτελούνται από δεδομένα εισόδου χωρίς επισημασμένες απαντήσεις. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετισμών και ομαδοποίησης.

Η ομαδοποίηση (clustering) είναι η πιο κοινή τεχνική μάθησης χωρίς επίβλεψη. Χρησιμοποιείται για διερευνητική ανάλυση δεδομένων για την εύρεση κρυφών μοτίβων ή ομαδοποιήσεων σε δεδομένα. Οι εφαρμογές για την ανάλυση συμπλέγματος (cluster analysis) περιλαμβάνουν ανάλυση αλληλουχίας γονιδίων, έρευνα αγοράς και αναγνώριση αντικειμένων.
Οι συνηθέστεροι αλγόριθμοι για την εκτέλεση συμπλεγμάτων περιλαμβάνουν k-means και k-medoids, ιεραρχική συσσώρευση (hierarchical clustering), μοντέλα Gaussian (Gaussian mixture models), μοντέλα Markov (Hidden Markov models), αυτό-οργανωτικούς χάρτες (self-organizing maps), ομαδοποίηση (clustering) και αφαιρετική συσσώρευση (subtractive clustering).

Στην **ενισχυτική μάθηση**, ένα πρόγραμμα υπολογιστή αλληλοεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας συγκεκριμένος στόχος. Το μηχάνημα εκπαιδεύεται να λαμβάνει συγκεκριμένες αποφάσεις. Το μηχάνημα εκτίθεται σε περιβάλλον όπου εκπαιδεύεται συνεχώς χρησιμοποιώντας δοκιμές και σφάλματα. Μαθαίνει από την εμπειρία του παρελθόντος και προσπαθεί να καταγράψει τις καλύτερες δυνατές γνώσεις για να λάβει ακριβείς αποφάσεις. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού.
Παράδειγμα της ενισχυτικής μάθησης είναι η διαδικασία απόφασης.

Κατηγοριοποίηση των προβλημάτων μηχανικής μάθησης, προκύπτει όταν κάποιος θεωρήσει το επιθυμητό αποτέλεσμα του συστήματος μηχανικής μάθησης :

- Κατά την ταξινόμηση (classification), τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή μάθηση πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Αυτό συνήθως εμπίπτει στην επιτηρούμενη μάθηση.
- Κατά την παλινδρόμηση (regression), επίσης πρόβλημα επιτηρούμενης μάθησης, τα αποτελέσματα είναι συνεχή και όχι διακριτά.
- Κατά την ομαδοποίηση (clustering), ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτόν τον διαχωρισμό τυπική εργασία μη επιτηρούμενης μάθησης.
- Η εκτίμηση πυκνότητας βρίσκει τη διανομή των δεδομένων εισόδου σε κάποιο χώρο.
- Σε προβλήματα μείωσης (dimensionality reduction), τα δεδομένα απλοποιούνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων. Το στατιστικό μοντέλο θεμάτων (Topic modeling) είναι ένα σχετικό πρόβλημα, όπου η μηχανή καλείται να βρει έγγραφα που καλύπτουν παρόμοια θέματα από ένα σύνολο εγγράφων γραμμένων σε φυσική γλώσσα.

## 3.2 Επιλογή αλγορίθμου

Η επιλογή του σωστού αλγορίθμου μπορεί να φανεί συντριπτική - υπάρχουν δεκάδες εποπτευόμενοι και ανεξέλεγκτοι αλγόριθμοι μηχανικής μάθησης και ο καθένας παίρνει μια διαφορετική προσέγγιση στη μάθηση.

Δεν υπάρχει καλύτερη μέθοδος ή ένα μέγεθος που να ταιριάζει σε όλα. Η εύρεση του σωστού αλγορίθμου είναι εν μέρει απλή δοκιμή και σφάλμα - ακόμη και οι έμπειροι ερευνητές δεδομένων δεν μπορούν να καταλάβουν εάν ένας αλγόριθμος θα λειτουργήσει χωρίς να το δοκιμάσουν. Αλλά η επιλογή αλγορίθμων εξαρτάται επίσης από το μέγεθος και τον τύπο των δεδομένων με τα οποία συνεργάζεστε, τις πληροφορίες που θέλετε να λάβετε από τα δεδομένα και τον τρόπο με τον οποίο θα χρησιμοποιηθούν αυτές οι πληροφορίες.

## **3.3 Αλγόριθμοι Μηχανικής Μάθησης**

### 3.3.1 Αλγόριθμοι Ταξινόμησης (Classification Algorithms)

Η ταξινόμηση είναι ένα από τα δεδομένα εξόρυξης. Αυτό χρησιμοποιείται για την ανάλυση ενός δεδομένου συνόλου δεδομένων και λαμβάνει κάθε εμφάνιση του. Αναθέτει αυτήν την εμφάνιση σε μια συγκεκριμένη κλάση. Τέτοιο ότι το σφάλμα κατάταξης θα είναι το λιγότερο. Χρησιμοποιείται για την εξαγωγή μοντέλων. Αυτό καθορίζει σημαντικές κατηγορίες δεδομένων εντός του δεδομένου συνόλου δεδομένων. Η ταξινόμηση είναι μια διαδικασία δύο σταδίων.

Κατά τη διάρκεια του πρώτου βήματος, το μοντέλο δημιουργείται εφαρμόζοντας αλγόριθμο ταξινόμησης. Αυτό συμβαίνει σε σύνολο δεδομένων εκπαίδευσης.

Στη συνέχεια, στο δεύτερο βήμα, το εξαγόμενο μοντέλο εξετάζεται σε σχέση με ένα προκαθορισμένο σύνολο δεδομένων δοκιμών. Αυτό είναι για να μετρήσει το εκπαιδευόμενο μοντέλο απόδοση και ακρίβεια. Έτσι, η ταξινόμηση είναι η διαδικασία για την εκχώρηση της ετικέτας κλάσης από ένα σύνολο δεδομένων του οποίου η ετικέτα κλάσης είναι άγνωστη.

Οι πιο δημοφιλείς αλγόριθμοι ταξινόμησης είναι:

- Classification and Regression Tree (CART)
- Support Vector Machines
- Naïve Bayes
- K-Nearest Neighbor(kNN)
- Random Forest
- K-Means
- Apriori
- Adaboost
- EM
- PageRank

### 3.3.2 Αλγόριθμοι Παλινδρόμησης (Regression Algorithms)

Η παλινδρόμηση ασχολείται με τη μοντελοποίηση της σχέσης μεταξύ των μεταβλητών, η οποία εκλέγεται με επαναληπτικό τρόπο χρησιμοποιώντας ένα μέτρο σφάλματος στις προβλέψεις του μοντέλου.

Οι μέθοδοι παλινδρόμησης είναι ένας άξονας των στατιστικών στοιχείων και έχουν συμμετάσχει στη στατιστική εκμάθηση μηχανών. Αυτό μπορεί να προκαλέσει σύγχυση επειδή μπορούμε να χρησιμοποιήσουμε την παλινδρόμηση για να αναφερθούμε στην κλάση του προβλήματος και την κλάση του αλγορίθμου.

Οι πιο δημοφιλείς αλγόριθμοι παλινδρόμησης είναι :

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

### 3.3.3 Αλγόριθμοι κατά περίπτωση (Instance-based Algorithms)

Το μοντέλο εκμάθησης βασισμένο σε περιστατικά είναι ένα πρόβλημα απόφασης με παραδείγματα ή παραδείγματα δεδομένων εκπαίδευσης που θεωρούνται σημαντικά ή απαιτούνται για το μοντέλο.

Τέτοιες μέθοδοι συνήθως δημιουργούν μια βάση δεδομένων με δεδομένα παραδείγματος και συγκρίνουν νέα δεδομένα με τη βάση δεδομένων χρησιμοποιώντας ένα μέτρο ομοιότητας για να βρουν την καλύτερη αντιστοιχία και να κάνουν μια πρόβλεψη. Για το λόγο αυτό, οι μέθοδοι βασισμένες σε στιγμές ονομάζονται επίσης μέθοδοι νίκης-λήψης όλων και μάθηση βασισμένη στη μνήμη. Εστιάζεται στην παρουσίαση των αποθηκευμένων ενδείξεων και των μέτρων ομοιότητας που χρησιμοποιούνται μεταξύ περιπτώσεων.

Οι πιο δημοφιλείς αλγόριθμοι κατά περίπτωση είναι:

- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)

- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

### 3.3.4 Αλγόριθμοι Τακτοποίησης (Regularization Algorithms)

Μια επέκταση σε μια άλλη μέθοδο (συνήθως μέθοδοι παλινδρόμησης) που επιβάλλουν τα μοντέλα με βάση την πολυπλοκότητά τους, ευνοώντας τα απλούστερα μοντέλα που είναι επίσης καλύτερα στη γενίκευση.

Οι πιο δημοφιλείς αλγόριθμοι τακτοποίησης είναι:

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

### 3.3.5 Αλγόριθμοι Δέντρων Αποφάσεων (Decision Tree Algorithms)

Οι μέθοδοι δέντρων αποφάσεων κατασκευάζουν ένα μοντέλο αποφάσεων που βασίζονται σε πραγματικές τιμές χαρακτηριστικών στα δεδομένα.

Τα δέντρα αποφάσεων είναι εκπαιδευμένα σε δεδομένα για προβλήματα ταξινόμησης και παλινδρόμησης. Τα δέντρα αποφάσεων είναι συχνά γρήγορα και ακριβή και είναι ένα μεγάλο φαβορί στη μηχανική μάθηση.

Οι πιο δημοφιλείς αλγόριθμοι δέντρων αποφάσεων είναι:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)

- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

### 3.3.6 Bayesian Algorithms

Οι Bayesian μέθοδοι είναι εκείνες που εφαρμόζουν ρητά το θεώρημα του Bayes για προβλήματα όπως ταξινόμηση και παλινδρόμηση.

Οι πιο δημοφιλείς Bayesian αλγόριθμοι είναι:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

### 3.3.7 Αλγόριθμοι Ομαδοποίησης (Clustering Algorithms)

Η ομαδοποίηση, όπως η παλινδρόμηση, περιγράφει την τάξη του προβλήματος και την κατηγορία των μεθόδων.

Οι μέθοδοι ομαδοποίησης οργανώνονται συνήθως από τις προσεγγίσεις μοντελοποίησης, όπως βασισμένη στο κέντρο και η ιεραρχική. Όλες οι μέθοδοι ασχολούνται με τη χρήση των εγγενών δομών στα δεδομένα για την καλύτερη οργάνωση των δεδομένων σε ομάδες μέγιστης κοινότητας.

Οι πιο δημοφιλείς αλγόριθμοι ομαδοποίησης είναι:

- k-Means

- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering

### 3.3.8 Αλγόριθμοι Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Network Algorithms)

Τα τεχνητά νευρωνικά δίκτυα είναι μοντέλα που εμπνέονται από τη δομή και / ή τη λειτουργία των βιολογικών νευρωνικών δικτύων.

Πρόκειται για μια τάξη αντιστοίχισης προτύπων που χρησιμοποιούνται συνήθως για προβλήματα παλινδρόμησης και ταξινόμησης, αλλά είναι πραγματικά ένας τεράστιος τομέας αποτελούμενος από εκατοντάδες αλγορίθμους και παραλλαγές για όλους τους τύπους προβλημάτων.

Οι πιο δημοφιλείς αλγόριθμοι τεχνητών νευρωνικών δικτύων είναι:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

# 4.Μηχανή Διανύσματος Υποστήριξης (Support Vector Machine) - Αλγόριθμος SVM

## 4.1 Εισαγωγή

Η μηχανή διανύσματος υποστήριξης (SVM) παρουσιάστηκε από τους Boser, Guyon και Vapnik. Αποτελούν ένα σύνολο συναφών μεθόδων μάθησης υπό επίβλεψη. Είναι μια προσέγγιση για την ταξινόμηση που αναπτύχθηκε στην κοινότητα των επιστημόνων πληροφορικής κατά τη δεκαετία του 1990 και έχει αυξηθεί σε δημοτικότητα από τότε. Ανήκουν σε μια οικογένεια γενικευμένων γραμμικών ταξινομητών. Η μηχανή διανύσματος υποστήριξης (SVM) είναι ένα εργαλείο πρόβλεψης, ταξινόμησης και παλινδρόμησης που χρησιμοποιεί τη θεωρία μάθησης μηχανών για τη μεγιστοποίηση της προβλεπτικής ακρίβειας, αποφεύγοντας αυτόματα την υπερβολική προσαρμογή στα δεδομένα και προέρχεται από την θεωρία της στατιστικής μάθησης[32].

Υποστηρικτικές μηχανές διανύσματος μπορούν να οριστούν ως συστήματα που χρησιμοποιούν υποθέσεις γραμμικών λειτουργιών σε ένα χώρο υψηλών διαστάσεων, εκπαιδευμένοι με έναν αλγόριθμο μάθησης από τη θεωρία βελτιστοποίησης που εφαρμόζει μια μαθησιακή προκατάληψη και προέρχεται από τη θεωρία της στατιστικής μάθησης. Ως μέθοδος ταξινόμησης, η SVM είναι ένα παγκόσμιο μοντέλο ταξινόμησης που δημιουργεί μη αλληλεπικαλυπτόμενα διαμερίσματα και συνήθως χρησιμοποιεί όλα τα χαρακτηριστικά. Ο χώρος της οντότητας χωρίζεται σε ένα μόνο πέρασμα, έτσι ώστε να παράγονται επίπεδα και γραμμικά διαμερίσματα. Οι SVM βασίζονται σε γραμμικά διακριτά περιθώρια μέγιστου περιθωρίου και είναι παρόμοια με τις προσεγγίσεις των πιθανοτήτων, αλλά δεν λαμβάνουν υπόψη τις εξαρτήσεις μεταξύ των χαρακτηριστικών[33].

Η μηχανή διανύσματος υποστήριξης ήταν αρχικά δημοφιλής στην κοινότητα NIPS και τώρα είναι ένα ενεργό μέρος της έρευνας μηχανών μάθησης σε όλο τον κόσμο. Το SVM γίνεται διάσημο όταν χρησιμοποιεί χάρτες pixel ως είσοδο. Δίνει ακρίβεια συγκρίσιμη με εξελιγμένα νευρωνικά δίκτυα με επεξεργασμένα χαρακτηριστικά σε μια εργασία αναγνώρισης χειρογράφου [34]. Παρόλο που θεωρείται ότι τα Νευρικά Δίκτυα είναι ευκολότερα χρήσιμα από αυτό, αποκτούνται μερικές φορές μη ικανοποιητικά αποτελέσματα. Ένα καθήκον ταξινόμησης συνήθως περιλαμβάνει δεδομένα κατάρτισης και δοκιμών που συνίστανται σε μερικές περιπτώσεις δεδομένων [35]. Κάθε παράσταση στο σετ κατάρτισης περιέχει έναν στόχο τιμές και πολλά χαρακτηριστικά. Ο στόχος του SVM είναι να παράγει ένα μοντέλο που προβλέπει την τιμή-στόχο των στιγμιότυπων δεδομένων στο σύνολο δοκιμών που δίδονται μόνο στα χαρακτηριστικά[36]. Χρησιμοποιείται επίσης για πολλές εφαρμογές, όπως ανάλυση χειρών, ανάλυση προσώπου και ούτω καθεξής, ειδικά για εφαρμογές ταξινόμησης προτύπων και

παλινδρόμησης. Τα θεμέλια των μηχανών διανύσματος υποστήριξης (SVM) έχουν αναπτυχθεί από τον Vapnik [37] και έχουν κερδίσει τη δημοτικότητα λόγω πολλών πολλά υποσχόμενων χαρακτηριστικών όπως η καλύτερη εμπειρική απόδοση. Η διατύπωση χρησιμοποιεί την αρχή της ελαχιστοποίησης των διαρθρωτικών κινδύνων (SRM), η οποία αποδείχθηκε ανώτερη [38] στην παραδοσιακή αρχή της Ελάχιστης Εμπειρικής Κινδύνου (ERM) που χρησιμοποιείται από τα συμβατικά νευρωνικά δίκτυα. Το SRM ελαχιστοποιεί ένα ανώτερο όριο στον αναμενόμενο κίνδυνο, όπου με το ERM ελαχιστοποιείται το σφάλμα στα δεδομένα εκπαίδευσης. Αυτή η διαφορά που παρέχει στην SVM μεγαλύτερη γενικευμένη ικανότητα, είναι ο στόχος της στατιστικής μάθησης. Τα SVM αναπτύχθηκαν για την επίλυση του προβλήματος ταξινόμησης, αλλά πρόσφατα έχουν επεκταθεί για την επίλυση προβλημάτων παλινδρόμησης [39].

Οι SVM έχουν αποδειχθεί ότι αποδίδουν καλά σε ποικίλες ρυθμίσεις και συχνά θεωρούνται ένας από τους καλύτερους ταξινομητές. Είναι μια γενίκευση ενός απλού και εντατικού ταξινομητή που ονομάζεται μέγιστος ταξινομητής περιθωρίου. Παρόλο που είναι κομψό και απλό, θα διαπιστώσουμε ότι αυτός ο ταξινομητής δυστυχώς δεν μπορεί να εφαρμοστεί στα περισσότερα σύνολα δεδομένων, καθώς απαιτεί το διαχωρισμό των κλάσεων από ένα γραμμικό όριο. Αντί να επιδιώκουμε το μεγαλύτερο περιθώριο έτσι ώστε κάθε παρατήρηση να μην είναι μόνο στη σωστή πλευρά του υπερπληθυσμού αλλά και στη σωστή πλευρά του περιθωρίου, επιτρέπουμε κάποιες παρατηρήσεις να βρίσκονται στην λανθασμένη πλευρά του περιθωρίου ή ακόμα και στην εσφαλμένη πλευρά του υπερπληθυσμού. (Το περιθώριο μπορεί να παραβιαστεί από κάποιες παρατηρήσεις εκπαίδευσης). Μια παρατήρηση μπορεί να είναι όχι μόνο στην λανθασμένη πλευρά του περιθωρίου, αλλά και στη λάθος πλευρά του υπερπληθυσμού. Στην πραγματικότητα, όταν δεν υπάρχει διαχωριστικό, μια τέτοια κατάσταση είναι αναπόφευκτη. Οι παρατηρήσεις σε λάθος πλευρά αντιστοιχούν σε παρατηρήσεις κατάρτισης που έχουν ταξινομηθεί εσφαλμένα από τον ταξινομητή διανύσματος υποστήριξης.

Ο ταξινομητής διανύσματος υποστήριξης ταξινομεί μια παρατήρηση δοκιμής ανάλογα με την πλευρά που βρίσκεται. Επιλέγεται για να διαχωρίσει σωστά τις περισσότερες από τις παρατηρήσεις εκπαίδευσης στις δύο κατηγορίες, αλλά μπορεί να ταξινομήσει εσφαλμένα μερικές παρατηρήσεις.

$$maximazeM \qquad (4.1)$$

$\beta_0,\beta_1,\ldots,\beta_p,e_1,\ldots,e_n,M$

subject to $\sum_{j=1}^{p} \beta^2 = 1,$ (4.2)

$y_i (\beta_0+\beta_{1xi1}+\ldots+\beta_{pxip}) \geq M(1-e_i)$ (4.3)

$e_i \geq 0, \sum_{i=1}^{n} ei \leq C,$ (4.4)

Είναι η λύση στο πρόβλημα βελτιστοποίησης όπου το C είναι μια παράμετρος μη αρνητικής ρύθμισης. Το M (4.1) είναι το πλάτος του περιθωρίου. Στο (4.3), ε1,. . . , εn είναι χαλαρές μεταβλητές που επιτρέπουν τις μεμονωμένες παρατηρήσεις να βρίσκονται στη λάθος πλευρά του περιθωρίου ή του υπερπληθυσμού. Το C (4.4) περιορίζει το άθροισμα των ei, και έτσι καθορίζει τον αριθμό και τη σοβαρότητα των παραβιάσεων στο περιθώριο. Μπορούμε να σκεφτούμε τον C ως προϋπολογισμό για το ποσό που το περιθώριο μπορεί να παραβιαστεί από τις παρατηρήσεις n. Αν το C = 0 τότε δεν υπάρχει προϋπολογισμός για παραβιάσεις στο περιθώριο, και πρέπει να ισχύει το γεγονός ότι e1 = ... = εn = 0. Για το C> 0 δεν μπορούν να εμφανιστούν περισσότερες από C παρατηρήσεις στην λανθασμένη πλευρά του υπερπληθυσμού, γιατί αν παρατηρηθεί μια λάθος πλευρά τότε ei> 1 και απαιτεί ότι ni = 1 ei ≤ C. Καθώς ο προϋπολογισμός Γ αυξάνεται, γινόμαστε πιο ανεκτικοί στις παραβιάσεις στο περιθώριο και έτσι το περιθώριο θα διευρυνθεί. Αντίθετα, καθώς η C μειώνεται, γινόμαστε λιγότερο ανεκτικοί στις παραβιάσεις στο περιθώριο και έτσι το περιθώριο στενεύει. Στην πράξη, το C αντιμετωπίζεται ως παράμετρος ρύθμισης που επιλέγεται γενικά μέσω διασταυρούμενης επικύρωσης. Το C ελέγχει το αντιστάθμισμα μεροληψίας-απόκλισης της στατιστικής μεθόδου μάθησης. Όταν το C είναι μικρό, αναζητούμε στενά περιθώρια που σπάνια παραβιάζονται. Αυτό ισοδυναμεί με έναν ταξινομητή που είναι πολύ κατάλληλος για τα δεδομένα, τα οποία μπορεί να έχουν χαμηλή μεροληψία αλλά μεγάλη διακύμανση. Από την άλλη πλευρά, όταν το C είναι μεγαλύτερο, το περιθώριο είναι ευρύτερο και επιτρέπουμε περισσότερες παραβιάσεις σε αυτό. Αυτό ισοδυναμεί με την προσαρμογή των δεδομένων λιγότερο σκληρά και την απόκτηση ενός ταξινομητή που είναι δυνητικά περισσότερο προκατειλημμένος, αλλά μπορεί να έχει μικρότερη διακύμανση. Το πρόβλημα βελτιστοποίησης (4.1) – (4.4) έχει μια πολύ ενδιαφέρουσα ιδιότητα και αποδεικνύει ότι μόνο οι παρατηρήσεις που είτε βρίσκονται στο περιθώριο είτε παραβιάζουν το περιθώριο θα επηρεάσουν τον ταξινομητή που αποκτήθηκε. Με άλλα λόγια, μια παρατήρηση που βρίσκεται ακριβώς στη σωστή πλευρά του περιθωρίου δεν επηρεάζει τον ταξινομητή φορέα υποστήριξης!

Η αλλαγή της θέσης αυτής της παρατήρησης δεν θα μεταβάλλει καθόλου τον ταξινομητή, υπό την προϋπόθεση ότι η θέση του παραμένει στη σωστή πλευρά του περιθωρίου. Οι παρατηρήσεις που βρίσκονται απευθείας στο περιθώριο ή στην λανθασμένη πλευρά του περιθωρίου για την τάξη τους, είναι γνωστές ως διανύσματα στήριξης. Αυτές οι παρατηρήσεις επηρεάζουν τον ταξινομητή διανύσματος υποστήριξης. Το γεγονός ότι μόνο τα διανύσματα υποστήριξης επηρεάζουν τον ταξινομητή είναι σύμφωνα με τον προηγούμενο ισχυρισμό μας ότι ο C ελέγχει την αντιστάθμιση της μεροληψίας-διακύμανσης του ταξινομητή διανύσματος υποστήριξης. Όταν η παράμετρος συντονισμού C είναι μεγάλη, τότε το περιθώριο είναι μεγάλο, πολλές παρατηρήσεις παραβιάζουν το περιθώριο και έτσι υπάρχουν πολλά διανύσματα υποστήριξης. Στην περίπτωση αυτή, πολλές παρατηρήσεις εμπλέκονται στον προσδιορισμό του υπερπληθυσμού. Αντίθετα, αν το C είναι μικρό, τότε θα υπάρχουν λιγότεροι φορείς υποστήριξης και επομένως ο ταξινομητής που θα προκύψει θα έχει χαμηλή μεροληψία αλλά μεγάλη διακύμανση. Το γεγονός ότι ο κανόνας απόφασης του ταξινομητή του διανύσματος υποστήριξης βασίζεται μόνο σε δυνητικά μικρό υποσύνολο των παρατηρήσεων εκπαίδευσης (τα διανύσματα υποστήριξης) σημαίνει ότι είναι αρκετά ανθεκτικοί στη συμπεριφορά παρατηρήσεων.

Η μηχανή διανύσματος υποστήριξης (SVM) είναι μια επέκταση του ταξινομητή φορέα υποστήριξης που προκύπτει από τη μεγέθυνση του χώρου χαρακτηριστικών με έναν συγκεκριμένο τρόπο, χρησιμοποιώντας πυρήνες.

Η υπερφόρτωση συχνά δεν φαίνεται να αποτελεί πρόβλημα, εν μέρει λόγω της ανυπαρξίας απώλειας εσφαλμένης ταξινόμησης.

Υπάρχουν πολλές διαφορετικές μέθοδοι για τη γενίκευση του κατηγοριοποιητή διανύσματος υποστήριξης δύο κατηγοριών σε κλάσεις K> 2.

- Στην προσέγγιση «one versus one», υπολογίζουμε όλους τους $\binom{K}{2}$ ταξινομητές ζεύγους. Για κάθε σημείο δοκιμής, η προβλεπόμενη κατηγορία είναι εκείνη που κερδίζει τους διαγωνισμούς με τα περισσότερα ζευγάρια.
- Στην προσέγγιση «one versus all», κάθε τάξη συγκρίνεται με όλες τις άλλες σε συγκρίσεις δύο κατηγοριών K. Για να ταξινομήσουμε ένα σημείο δοκιμής, υπολογίζουμε την εμπιστευτικότητα για κάθε έναν από τους ταξινομητές K. Ο νικητής είναι η τάξη με την υψηλότερη εμπιστοσύνη.

Τέλος, οι Vapnik (1998) και οι Weston και Watkins (1999) πρότειναν (κάπως πολύπλοκα) κριτήρια πολλαπλών κατηγοριών που γενικεύουν το κριτήριο δύο κατηγοριών. Οι Tibshirani και

Hastie (2007) προτείνουν τον ταξινομητή δένδρων περιθωρίου, στον οποίο χρησιμοποιούνται ταξινομητές διανύσματος υποστήριξης σε ένα δυαδικό δέντρο, όπως και στο CART. Οι τάξεις οργανώνονται με ιεραρχικό τρόπο, οι οποίες μπορεί να είναι χρήσιμες για την ταξινόμηση των ασθενών σε διαφορετικούς τύπους καρκίνου, για παράδειγμα[40].

Όταν τα SVM εισήχθησαν για πρώτη φορά στα μέσα της δεκαετίας του 1990, έκαναν εκτόξευση στις κοινότητες στατιστικής και μηχανικής μάθησης. Η ιδέα της εξεύρεσης υπερπληθυσμού που να διαχωρίζει τα δεδομένα όσο το δυνατόν περισσότερο, ενώ ελαχιστοποιεί κάποιες παραβιάσεις σε αυτό το διαχωρισμό, φαινόταν σαφώς διαφορετική από τις κλασσικές προσεγγίσεις ταξινόμησης, όπως η λογιστική παλινδρόμηση και η ανάλυση διακρίσεων γραμμών. Επιπλέον, η ιδέα της χρησιμοποίησης ενός πυρήνα για την επέκταση του χώρου χαρακτηριστικών για την προσαρμογή των μη γραμμικών ορίων τάξης φάνηκε να είναι ένα μοναδικό και πολύτιμο χαρακτηριστικό.

Ωστόσο, από τότε, έχουν προκύψει βαθιές συνδέσεις μεταξύ SVM και άλλων πιο κλασσικών στατιστικών μεθόδων. Έχουμε διαπιστώσει ότι ο ταξινομητής διανύσματος υποστήριξης συνδέεται στενά με την υλικοτεχνική παλινδρόμηση και άλλες προ υπάρχουσες στατιστικές μεθόδους. Ωστόσο, για ιστορικούς λόγους, η χρήση μη γραμμικών πυρήνων είναι πολύ πιο διαδεδομένη στο πλαίσιο των SVM από ό, τι στο πλαίσιο της λογιστικής παλινδρόμησης ή άλλων μεθόδων[41].

### 4.2 Αλγόριθμος

Η SVM είναι ένας αλγόριθμος για την εκμάθηση μισών χώρων με έναν ορισμένο τύπο προηγούμενης γνώσης, δηλαδή την προτίμηση για μεγάλο περιθώριο. Το hard-SVM επιδιώκει το μισό διάστημα που χωρίζει τέλεια τα δεδομένα με το μεγαλύτερο περιθώριο, ενώ το soft-SVM δεν αναλαμβάνει τη δυνατότητα διαχωρισμού των δεδομένων και επιτρέπει να παραβιαστούν οι περιορισμοί σε κάποιο βαθμό. Η πολυπλοκότητα του δείγματος και για τους δύο τύπους SVM είναι διαφορετική από την πολυπλοκότητα του δείγματος της απλής διδασκαλίας του μισού χώρου[42].

Η επίτευξη καλών αποτελεσμάτων με μηχανές διανύσματος υποστήριξης απαιτεί κάποια προσοχή. Η συναινετική άποψη είναι ότι ο καλύτερος ταξινομητής SVM είναι συνήθως

τουλάχιστον τόσο ακριβής όσο ο καλύτερος από οποιοδήποτε άλλο τύπο ταξινομητή, αλλά η απόκτηση του καλύτερου ταξινομητή SVM δεν είναι ασήμαντη.

Η ακόλουθη διαδικασία συνιστάται ως σημείο εκκίνησης:

1. Κωδικοποίηση των δεδομένων σε αριθμητική μορφή που απαιτείται για την εκπαίδευση SVM.

2. Κλιμάκωση κάθε χαρακτηριστικού να έχει εύρος 0 έως 1, ή να έχει εύρος -1 έως +1, ή να έχει μέση διαφορά μηδέν και μονάδας.

3. Χρήση της πολλαπλής επικύρωσης για να βρεθεί η καλύτερη τιμή για το C για έναν γραμμικό πυρήνα.

4. Χρήση διασταυρούμενης επικύρωσης για να βρεθούν οι καλύτερες τιμές για C και γ για έναν πυρήνα RBF.

5. Εκπαίδευση όλων των διαθέσιμων δεδομένα χρησιμοποιώντας τις τιμές των παραμέτρων που βρέθηκαν να είναι καλύτερες μέσω διασταυρούμενης επικύρωσης[43].

Η ταξινόμηση στο SVM είναι ένα παράδειγμα εποπτευόμενης μάθησης. Οι γνωστές ετικέτες υποδεικνύουν εάν το σύστημα εκτελεί σωστά ή όχι. Αυτές οι πληροφορίες δείχνουν την επιθυμητή απόκριση, επικυρώνουν την ακρίβεια του συστήματος ή χρησιμοποιούνται για να βοηθήσουν το σύστημα να μάθει να ενεργεί σωστά. Ένα βήμα στην ταξινόμηση SVM περιλαμβάνει την ταυτοποίηση που είναι στενά συνδεδεμένη με τις γνωστές τάξεις. Αυτό ονομάζεται επιλογή χαρακτηριστικών ή εξαγωγή χαρακτηριστικών. Η επιλογή χαρακτηριστικών και η ταξινόμηση SVM από κοινού έχουν μια χρήση ακόμη και όταν δεν είναι απαραίτητη η πρόβλεψη άγνωστων δειγμάτων. Μπορούν να χρησιμοποιηθούν για τον προσδιορισμό των συνόλων κλειδιών που εμπλέκονται σε οποιεσδήποτε διαδικασίες διακρίνουν τις κλάσεις [36].

Το μοντέλο μηχανής διανύσματος υποστήριξης (SVM) για την εκμάθηση γραμμικών προτύπων σε χώρους υψηλών διαστάσεων. Η μεγάλη διάσταση του χώρου χαρακτηριστικών

αυξάνει τόσο την πολυπλοκότητα του δείγματος όσο και τις προκλήσεις της υπολογιστικής πολυπλοκότητας.

Το αλγοριθμικό μοντέλο SVM αντιμετωπίζει την πρόκληση πολυπλοκότητας του δείγματος αναζητώντας διαχωριστές μεγάλου περιθωρίου. Συνοπτικά, ένας μισός χώρος χωρίζει ένα σετ εκπαίδευσης με μεγάλο περιθώριο αν όλα τα παραδείγματα δεν είναι μόνο στη σωστή πλευρά του διαχωριστικού υπερπληρωμένου αλλά και πολύ μακριά από αυτό. Ο περιορισμός του αλγορίθμου για την παραγωγή ενός μεγάλου διαχωριστή περιθωρίου μπορεί να αποδώσει μια μικρή πολυπλοκότητα δείγματος, ακόμη και αν η διασταλτικότητα του χώρου χαρακτηριστικών είναι υψηλή (και μάλιστα απεριόριστη).

**Πλεονεκτήματα**

- Οι SVM μπορούν να μοντελοποιήσουν μη γραμμικά φαινόμενα με την επιλογή μιας κατάλληλης μεθόδου πυρήνα.
- Οι SVM παρέχουν γενικά ακριβείς προβλέψεις.
- Οι SVM καθορίζουν το βέλτιστο διαχωριστικό του υπερπληθυσμού από τα πλησιέστερα σημεία (διανύσματα υποστήριξης) και όχι από τα απομακρυσμένα σημεία. Αυτό ενισχύει έτσι την ευρωστία του μοντέλου σε ορισμένες περιπτώσεις.

**Μειονέκτημα**

- Τα μοντέλα είναι αδιαφανή.
- Παρόλο που μπορούν να εξηγηθούν με ένα δέντρο αποφάσεων, υπάρχει κίνδυνος απώλειας ή ακρίβειας.
- Οι SVM είναι πολύ ευαίσθητες στην επιλογή των παραμέτρων του πυρήνα. Η δυσκολία στην επιλογή των σωστών παραμέτρων του πυρήνα μπορεί να σας υποχρεώσει να δοκιμάσετε πολλές πιθανές τιμές.
- Ως αποτέλεσμα, ο χρόνος υπολογισμού είναι μερικές φορές μακρύς[44].

# 5. Αλγόριθμος Naïve Bayes

## 5.1 Εισαγωγή

Η Bayesian λογική παρέχει μια πιθανότητα προσέγγισης στην εξαγωγή συμπερασμάτων. Βασίζεται στην υπόθεση ότι οι ποσότητες ενδιαφέροντος διέπονται από κατανομές πιθανοτήτων και ότι οι βέλτιστες αποφάσεις μπορούν να ληφθούν με τη συλλογιστική σχετικά με αυτές τις δυνατότητες μαζί με τα παρατηρούμενα δεδομένα. Είναι σημαντικό να μάθει η μηχανική μάθηση, επειδή παρέχει μια ποσοτική προσέγγιση για τη στάθμιση των αποδεικτικών στοιχείων που υποστηρίζουν εναλλακτικές υποθέσεις. Η Bayesian λογική παρέχει τη βάση για αλγορίθμους μάθησης που χειρίζονται άμεσα τις πιθανότητες, καθώς και ένα πλαίσιο για την ανάλυση της λειτουργίας άλλων αλγορίθμων που δεν χειρίζονται ρητά τις πιθανότητες[45].

Οι μέθοδοι μάθησης Bayesian σχετίζονται με τη μελέτη της μηχανικής μάθησης για δύο διαφορετικούς λόγους. Πρώτον, οι Bayesian αλγόριθμοι μάθησης που υπολογίζουν σαφείς πιθανότητες για υποθέσεις, όπως ο ταξινομητής Bayes, είναι μεταξύ των πιο πρακτικών προσεγγίσεων σε ορισμένα είδη μαθησιακών προβλημάτων. Για παράδειγμα, οι Michie et al. (1994) παρέχουν μια λεπτομερή μελέτη που συγκρίνει τον ταξινομητή Bayes σε άλλους αλγορίθμους μάθησης, συμπεριλαμβανομένων αλγορίθμων αποφάσεων και αλγορίθμων νευρωνικών δικτύων. Αυτοί οι ερευνητές δείχνουν ότι ο ταξινομητής Bayes είναι ανταγωνιστικός με αυτούς τους άλλους αλγορίθμους μάθησης σε πολλές περιπτώσεις και ότι σε ορισμένες περιπτώσεις υπερέχει αυτών των άλλων μεθόδων. Ο ταξινομητής Bayes είναι ένας από τους πιο αποτελεσματικούς αλγόριθμους που είναι γνωστοί. Ο δεύτερος λόγος που οι μέθοδοι Bayesian είναι σημαντικοί για τη μελέτη της μηχανικής μάθησης είναι ότι παρέχουν μια χρήσιμη προοπτική για την κατανόηση πολλών αλγορίθμων μάθησης που δεν χειρίζονται ρητά τις πιθανότητες.

Χρησιμοποιούμε επίσης μια Bayesian ανάλυση για να δικαιολογήσουμε μια βασική επιλογή σχεδιασμού στη μάθηση νευρωνικών δικτύων και στόχος είναι να ελαχιστοποιηθεί το άθροισμα των τετραγωνικών σφαλμάτων κατά την αναζήτηση πιθανών νευρωνικών δικτύων. Παίρνουμε επίσης μια εναλλακτική λειτουργία σφάλματος, cross entropy, το οποίο είναι πιο κατάλληλο από το άθροισμα των τετραγωνικών σφαλμάτων κατά την εκμάθηση λειτουργιών

στόχων που προβλέπουν πιθανότητες. Χρησιμοποιούμε μια Bayesian προοπτική για να αναλύσει την επαγωγική προκατάληψη των αλγορίθμων μάθησης δέντρων αποφάσεων που ευνοούν το σύντομο δέντρο αποφάσεων και να εξετάσει την αρχή που σχετίζεται στενά με το ελάχιστο μήκος. Μια βασική εξοικείωση με τις Bayesian μεθόδους είναι σημαντική για την κατανόηση και χαρακτηρίζει τη λειτουργία πολλών αλγορίθμων στη μηχανική μάθηση.

**Χαρακτηριστικά των μεθόδων μάθησης Bayesian:**

- Κάθε παρατηρούμενο παράδειγμα εκπαίδευσης μπορεί να μειώσει ή να αυξήσει την εκτιμώμενη πιθανότητα ότι μια υπόθεση είναι σωστή. Αυτό παρέχει μια πιο ευέλικτη προσέγγιση στη μάθηση από αλγορίθμους που εξαλείφουν εντελώς μια υπόθεση εάν διαπιστωθεί ότι είναι ασυμβίβαστη με οποιοδήποτε παράδειγμα. Οι προηγούμενες γνώσεις μπορούν να συνδυαστούν με τα παρατηρούμενα δεδομένα για να καθοριστεί η τελική πιθανότητα για μια υπόθεση.

- Στην Bayesian μάθηση, οι προηγούμενες γνώσεις προσφέρονται με την επιβεβαίωση (1) μίας προηγούμενης πιθανότητας για κάθε υποψήφια υπόθεση και (2) μια κατανομή πιθανότητας πάνω από τα παρατηρούμενα δεδομένα για κάθε πιθανή υπόθεση.

- Οι Bayesian μέθοδοι μπορούν να φιλοξενήσουν υποθέσεις που κάνουν πιθανότητες πρόβλεψης (π.χ. υποθέσεις όπως "αυτός ο ασθενής πνευμονίας έχει 93% πιθανότητα πλήρους ανάκαμψης"). Νέες καταστάσεις μπορούν να ταξινομηθούν συνδυάζοντας τις προβλέψεις πολλών υποθέσεων, σταθμισμένες από τις πιθανότητές τους. Ακόμα και στις περιπτώσεις όπου οι Bayesian μέθοδοι αποδειχθούν υπολογιστικές δυσκολίες, μπορούν να παράσχουν ένα πρότυπο βέλτιστης λήψης αποφάσεων έναντι του οποίου μπορούν να μετρηθούν και άλλες πρακτικές μέθοδοι.

- Μια πρακτική δυσκολία στην εφαρμογή Bayesian μεθόδων είναι ότι συνήθως απαιτούν την αρχική γνώση πολλών πιθανοτήτων. Όταν αυτές οι πιθανότητες δεν είναι γνωστές εκ των προτέρων, συχνά εκτιμώνται με βάση τις βασικές γνώσεις, τα προηγουμένως

διαθέσιμα δεδομένα και τις υποθέσεις σχετικά με τη μορφή των υποκείμενων κατανομών.

- Μια δεύτερη πρακτική δυσκολία είναι το σημαντικό υπολογιστικό κόστος που απαιτείται για τον προσδιορισμό της βέλτιστης υποθέσεως Bayes στη γενική περίπτωση (γραμμική στον αριθμό των υποψήφιων υποθέσεων). Σε ορισμένες εξειδικευμένες καταστάσεις, αυτό το υπολογιστικό κόστος μπορεί να μειωθεί σημαντικά.

## 5.2 BAYES THEOREM

Το θεώρημα του Bayes πήρε το όνομά του από τον Thomas Bayes, έναν Άγγλο κληρικό, ο οποίος έκανε πρώιμη εργασία στην θεωρία της πιθανότητας και της απόφασης κατά τη διάρκεια του 18ου αιώνα. Ας υποθέσουμε ότι X είναι μια πλειάδα δεδομένων. Στους Bayesian όρους, το X θεωρείται ως «αποδεικτικό στοιχείο». Ως συνήθως, περιγράφεται με μετρήσεις που έγιναν σε ένα σύνολο χαρακτηριστικών n. Έστω h κάποια υπόθεση όπως το ότι η πλειάδα δεδομένων X ανήκει σε μια συγκεκριμένη κατηγορία C. Για προβλήματα ταξινόμησης, θέλουμε να καθορίσουμε την P (H | X), την πιθανότητα ότι η υπόθεση H δίνεται η «απόδειξη» ή παρατηρούνται πλειάδα δεδομένων X. Με άλλα λόγια, ψάχνουμε για την πιθανότητα η τάξη X να ανήκει στην τάξη C, δεδομένου ότι γνωρίζουμε την περιγραφή χαρακτηριστικών του X. P (H | X) είναι η οπίσθια πιθανότητα, ή εκ των υστέρων πιθανότητας, των H στο X.

*Υπολογισμός πιθανότητας:*

P (H), P (X | H) και P (X) μπορεί να υπολογιστεί από τα δεδομένα. Το θεώρημα του Bayes είναι χρήσιμο στο ότι παρέχει έναν τρόπο υπολογισμού της οπίσθιας πιθανότητας, P (H | X), από P (H), P (X | H) και P (X). Το θεώρημα του Bayes είναι P (H | X) = P (X | H) P (H).

**Το θεώρημα Bayes** παρέχει έναν τρόπο υπολογισμού της πιθανότητας μιας υποθέσης με βάση την προηγούμενη πιθανότητα, τις πιθανότητες παρατήρησης διαφόρων δεδομένων που έχουν δοθεί στην υπόθεση και τα ίδια τα δεδομένα που παρατηρούνται. Οι Bayesian μέθοδοι επιτρέπουν την εκχώρηση μιας πιθανότητας posterior σε κάθε υποψήφια υπόθεση, με βάση

αυτούς τους υποτιθέμενους priors και τα παρατηρούμενα δεδομένα. Μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της πιο πιθανής υπόθεσης δεδομένων των δεδομένων - της μέγιστης a posteriori (MAP) υπόθεσης. Αυτή είναι η βέλτιστη υπόθεση υπό την έννοια ότι καμία άλλη υπόθεση δεν είναι πιο πιθανή. Για να ορίσουμε με ακρίβεια το θεώρημα Bayes, ας εισαγάγουμε πρώτα μια μικρή συμβολική αναφορά. Θα γράψουμε το P (h) για να υποδηλώσουμε την αρχική πιθανότητα της υποθέσεως h, πριν παρατηρήσουμε τα δεδομένα εκπαίδευσης. Το P (h) ονομάζεται συχνά η προηγούμενη πιθανότητα του h και μπορεί να αντανακλά οποιαδήποτε γνώση του υποβάθρου που έχουμε για την πιθανότητα ότι το h είναι μια σωστή υπόθεση. Εάν δεν έχουμε τέτοια προηγούμενη γνώση, τότε ίσως απλά αναθέσουμε την ίδια προηγούμενη πιθανότητα σε κάθε υποψήφια υπόθεση. Ομοίως, θα γράψουμε το P (D) για να υποδηλώσουμε την προηγούμενη πιθανότητα ότι τα δεδομένα εκπαίδευσης D θα παρατηρηθούν (δηλαδή, η πιθανότητα D να μην γνωρίζει ποια υπόθεση ισχύει). Στη συνέχεια, θα γράψουμε το P (D1h) για να υποδείξουμε την πιθανότητα παρατήρησης των δεδομένων D που δίνεται σε κάποιο κόσμο στον οποίο ισχύει η υπόθεση h. Γενικότερα, γράφουμε P (xly) για να δηλώσουμε την πιθανότητα του x που δίνεται στο y. Σε προβλήματα μηχανικής μάθησης ενδιαφέρεστε για την πιθανότητα P (h1D) που κρατάει το h δεδομένου των παρατηρούμενων δεδομένων εκπαίδευσης D. Το P (h1D) ονομάζεται posterior probability (η στατιστική πιθανότητα ότι μια υπόθεση είναι αληθής υπολογίζεται με βάση τις σχετικές παρατηρήσεις) της h, διότι αντανακλά την εμπιστοσύνη που έχουμε στην h αφού έχουμε δει τα δεδομένα της εκπαίδευσης D. Παρατηρήστε ότι η οπίσθια πιθανότητα P (h1D) αντικατοπτρίζει την επίδραση των δεδομένων εκπαίδευσης D, σε αντίθεση με την προηγούμενη η πιθανότητα P (h), η οποία είναι ανεξάρτητη από τον D.

### 5.3 Ταξινομητής BAYES

Οι Bayesian ταξινομητές είναι στατιστικοί ταξινομητές. Μπορούν να προβλέψουν πιθανότητες συμμετοχής στην τάξη, όπως η πιθανότητα να ανήκει μια δεδομένη πλειάδα σε μια συγκεκριμένη κλάση. Η Bayesian ταξινόμηση βασίζεται στο θεώρημα του Bayes. Οι μελέτες σύγκρισης των αλγορίθμων ταξινόμησης έχουν βρει έναν απλό Bayesian ταξινομητή γνωστό ως ο Naïve Bayes classifier να είναι συγκρίσιμος σε απόδοση με το δέντρο αποφάσεων και τους

επιλεγμένους ταξινομητές νευρωνικών δικτύων. Οι Bayesian ταξινομητές έχουν επίσης επιδείξει υψηλή ακρίβεια και ταχύτητα όταν εφαρμόζονται σε μεγάλες βάσεις δεδομένων. Οι Naive Bayesian ταξινομητές υποθέτουν ότι η επίδραση μιας τιμής χαρακτηριστικού σε μια δεδομένη κλάση είναι ανεξάρτητη από τις τιμές των άλλων χαρακτηριστικών. Αυτή η υπόθεση ονομάζεται ανεξαρτησία κατά κατηγορία. Έχει σκοπό να απλοποιήσει τους σχετικούς υπολογισμούς και, με αυτή την έννοια, θεωρείται "naive."

Ο ταξινομητής Naïve Bayes εφαρμόζεται σε μαθησιακές εργασίες όπου κάθε εμφάνιση x περιγράφεται από συνδυασμό τιμών αξιών και όπου η συνάρτηση στόχος f (x) μπορεί να πάρει οποιαδήποτε τιμή από κάποια πεπερασμένη ομάδα V. Μια ενδιαφέρουσα διαφορά μεταξύ της μεθόδου εκμάθησης του Bayes και των άλλων μεθόδων εκμάθησης είναι ότι δεν υπάρχει ρητή αναζήτηση μέσα από το χώρο των πιθανών υποθέσεων. Αντ' αυτού, η υπόθεση δημιουργείται χωρίς αναζήτηση, απλά μετρώντας τη συχνότητα των διαφόρων συνδυασμών δεδομένων μέσα στα παραδείγματα εκπαίδευσης.

Πολλές από τις βασικές έννοιες των Bayesian ταξινομητών και των λιγότερο τετράγωνων ταξινομητών σφαλμάτων συζητούνται από τους Duda και Hart (1973). Ο Domingos και ο Pazzani (1996) παρέχουν μια ανάλυση των συνθηκών κάτω από τις οποίες οι naive Bayes θα παράγουν βέλτιστες ταξινομήσεις, ακόμη και όταν παραβιάζεται η παραδοχή της ανεξαρτησίας τους (το κλειδί εδώ είναι ότι υπάρχουν συνθήκες κάτω από τις οποίες θα προκύψουν βέλτιστες ταξινομήσεις ακόμα και όταν οι σχετικές εκτιμήσεις posterior probabilities είναι εσφαλμένα). Ο Cestnik (1990) παρέχει μια συζήτηση σχετικά με τη χρήση της εκτίμησης m για την εκτίμηση πιθανών. Πειραματικά αποτελέσματα που συγκρίνουν διάφορες Bayesian προσεγγίσεις για τη μάθηση δέντρων αποφάσεων και άλλοι αλγόριθμοι μπορούν να βρεθούν στο Michie et al. (1994). Μια συζήτηση για την αρχή του ελάχιστου μήκους περιγραφής μπορεί να βρεθεί στο Rissanen (1983, 1989). Οι Quinlan και Rivest (1989) περιγράφουν τη χρήση τους στην αποφυγή υπερφόρτωσης στα δέντρα αποφάσεων[30].

***Ο Bayesian ταξινομητής λειτουργεί ως εξής:***

1. D είναι ένα σύνολο εκπαιδευτικών πλειάδων και των σχετικών ετικετών κλάσης τους. Ως συνήθως, κάθε πλειάδα αντιπροσωπεύεται από ένα δισδιάστατο διάνυσμα

χαρακτηριστικών, X = (χ1, χ2, ..., χν), που απεικονίζει n μετρήσεις που έγιναν στην πλειάδα από τα χαρακτηριστικά n, αντιστοίχως A1, A2, ....

2. Υποθέστε ότι υπάρχουν τάξεις m, C1, C2, ..., Cm. Με δεδομένη την πλειάδα, X, ο ταξινομητής θα προβλέψει ότι το X ανήκει στην τάξη που έχει την υψηλότερη posterior πιθανότητα, που εξαρτάται από το X. Δηλαδή, ο παριστάμενος Bayesian ταξινομητής προβλέπει ότι η πλειάδα X ανήκει στην κλάση Ci και μόνο αν P (C 1 | X)> P (Cj | X) για 1≤j≤m, j = i. Έτσι μεγιστοποιούμε το P (Ci | X). Προσδιορίζεται η κλάση Ci για την οποία μεγιστοποιείται το P (Ci | X) η μέγιστη εκ των υστέρων υπόθεση. Με το θεώρημα του Bayes, P (Ci | X) = P (X | O) P (01). P (X) Μέθοδοι ταξινόμησης Bayes.

3. Καθώς το P (X) είναι σταθερό για όλες τις κλάσεις, μόνο το P (X | Ci) P (Ci) πρέπει να μεγιστοποιηθεί. Αν οι προηγούμενες πιθανότητες δεν είναι γνωστές, τότε συνήθως θεωρείται ότι οι κλάσεις είναι εξίσου πιθανές, δηλαδή, P (C1) = P (C2) = ⋯ = P (Cm) X | Ci). Διαφορετικά μεγιστοποιούμε το P (X | Ci) P (Ci). Σημειώστε ότι οι προγενέστερες πιθανότητες κατηγορίας μπορούν να εκτιμηθούν με P (Ci) = | Ci, D | / | D |, όπου | Ci, D | είναι ο αριθμός των εκπαιδευτικών πλειάδων κατηγορίας Ci στο D.

4. Δεδομένων των συνόλων δεδομένων με πολλά χαρακτηριστικά, θα ήταν εξαιρετικά δαπανηρό να υπολογιστεί το P (X | Ci). Για να μειώσουμε τον υπολογισμό στην αξιολόγηση του P (X | Ci), γίνεται η παραδοχή της ανεξαρτησίας της τάξης υπό όρους. Αυτό προϋποθέτει ότι οι τιμές των χαρακτηριστικών είναι ανεξάρτητα από την μία από την άλλη, δεδομένης της ετικέτας της κλάσης της πλειάδας (δηλ. Ότι δεν υπάρχουν σχέσεις εξάρτησης μεταξύ των χαρακτηριστικών). Ετσι, n P (X | Ci) = P (xk | Ci) k = 1 = P (x1 | Ci) × P (x2 | Ci) × ⋯ × P (xn | Ci). Μπορούμε εύκολα να υπολογίσουμε τις πιθανότητες P (x1 | Ci), P (x2 | Ci), ..., P (xn | Ci) από τις πλειάδες εκπαίδευσης. Υπενθυμίζουμε ότι εδώ το xk αναφέρεται στην τιμή του χαρακτηριστικού Ak για την πλειάδα X. Για κάθε χαρακτηριστικό, εξετάζουμε εάν το χαρακτηριστικό είναι κατηγορηματικό ή συνεχές. Για παράδειγμα, για τον υπολογισμό του P (X | Ci), θεωρούμε τα εξής: (a) Αν το Ak είναι κατηγορηματικό, τότε P (xk | Ci) είναι ο αριθμός πλειάδων της κλάσης Ci στο D που έχει την τιμή xk για Ak, διαιρούμενη με | Ci, D |, τον

αριθμό πλειάδων της κατηγορίας Ci στο D. (β) Εάν το Ak είναι συνεχής, τότε πρέπει να κάνουμε λίγο περισσότερη δουλειά, αλλά ο υπολογισμός είναι αρκετά απλός. Ένα χαρακτηριστικό συνεχούς αποτίμησης θεωρείται ότι έχει Gaussian κατανομή με μέση μ και τυπική απόκλιση σ, που ορίζεται από g (x, μ, σ) = √ 1 - (χ-μ) 2 e 2σ2, έτσι ώστε 2πσ P (xk | Ci) = g (xk, μCi, σCi). Πρέπει να υπολογίσουμε τα μCi και σCi, τα οποία είναι ο μέσος όρος και η τυπική απόκλιση, αντίστοιχα, των τιμών του χαρακτηριστικού Ak για τις πλειάδες κατάρτισης της κλάσης Ci. Στη συνέχεια, συνδέουμε αυτές τις δύο ποσότητες σε εξίσωση, μαζί με το xk, για να υπολογίσουμε το P (xk | Ci).

*Όσων αφορά την αποτελεσματικότητα των Bayesian ταξινομητών:*

Διάφορες εμπειρικές μελέτες αυτού του ταξινομητή σε σύγκριση με το δέντρο αποφάσεων και τους ταξινομητές νευρωνικών δικτύων το έχουν βρει να είναι συγκρίσιμος σε ορισμένους τομείς. Θεωρητικά, οι Bayesian ταξινομητές έχουν το ελάχιστο ποσοστό σφάλματος σε σύγκριση με όλους τους άλλους ταξινομητές. Εντούτοις, στην πράξη αυτό δεν συμβαίνει πάντοτε, λόγω ανακριβειών στις παραδοχές που έγιναν για τη χρήση του, όπως η ανεξαρτησία κατά κατηγορία και η έλλειψη διαθέσιμων δεδομένων πιθανότητας. Οι Bayesian ταξινομητές είναι επίσης χρήσιμοι επειδή παρέχουν μια θεωρητική δικαιολογία για άλλους ταξινομητές που δεν χρησιμοποιούν ρητά το θεώρημα του Bayes. Για παράδειγμα, κάτω από ορισμένες υποθέσεις, μπορεί να αποδειχθεί ότι πολλοί αλγόριθμοι νευρωνικού δικτύου και προσαρμογής καμπυλών εξάγουν την υπόθεση της μέγιστης εκ των υστέρων, όπως και ο αρχικός Bayesian ταξινομητής[46].

Στη μηχανική μάθηση ενδιαφερόμαστε συχνά να καθορίσουμε την καλύτερη υπόθεση από κάποιο χώρο H, λαμβάνοντας υπόψη τα παρατηρούμενα δεδομένα εκπαίδευσης D. Ένας τρόπος να προσδιορίσουμε τι εννοούμε με την καλύτερη υπόθεση είναι να πούμε ότι απαιτούμε την πιο πιθανή υπόθεση, των δεδομένων D καθώς και οποιαδήποτε αρχική γνώση σχετικά με τις προηγούμενες πιθανότητες των διαφόρων υποθέσεων στο θεώρημα H. Το θεώρημα Bayes παρέχει μια άμεση μέθοδο για τον υπολογισμό τέτοιων πιθανοτήτων.

***Πλεονεκτήματα***

- Είναι εύκολο και γρήγορο να προβλεφθεί η τάξη των συνόλων δοκιμών.
- Είναι αποτελεσματικό στην πρόβλεψη πολλαπλών τάξεων.
- Όταν την παραδοχή της ανεξαρτησίας κατέχει ένας ταξινομητής Naive Bayes εκτελεί καλύτερη σύγκριση με άλλα μοντέλα, όπως η λογιστική παλινδρόμηση και χρειάζονται λιγότερα δεδομένα εκπαίδευσης.
- Λειτουργεί καλά σε περίπτωση κατηγορικών μεταβλητών εισόδου σε σύγκριση με αριθμητικές μεταβλητές. Για αριθμητική μεταβλητή, θεωρείται κανονική κατανομή (καμπύλη καμπάνας, η οποία είναι ισχυρή παραδοχή).

***Μειονεκτήματα***

- Εάν η κατηγοριοποιητική μεταβλητή έχει μια κατηγορία (σε σύνολο δεδομένων δοκιμής), η οποία δεν παρατηρήθηκε σε σύνολο δεδομένων εκπαίδευσης, τότε το μοντέλο θα αποδώσει μια πιθανότητα 0 (μηδέν) και δεν θα μπορέσει να κάνει πρόβλεψη. Αυτό είναι συχνά γνωστό ως "μηδενική συχνότητα". Για να το λύσουμε αυτό, μπορούμε να χρησιμοποιήσουμε την τεχνική εξομάλυνσης. Μία από τις πιο απλές τεχνικές εξομάλυνσης ονομάζεται εκτίμηση Laplace.
- Από την άλλη πλευρά, ο Naive Bayes είναι επίσης γνωστός ως κακός εκτιμητής, επομένως τα πιθανά αποτελέσματα από το predict_proba δεν πρέπει να ληφθούν πολύ σοβαρά υπόψη.
- Ένας άλλος περιορισμός των Naive Bayes είναι η υπόθεση των ανεξάρτητων προγνωστικών. Στην πραγματική ζωή, είναι σχεδόν αδύνατο να έχουμε ένα σύνολο προγνωστικών που είναι εντελώς ανεξάρτητα.

## 6. Μάθηση Δέντρων Αποφάσεων (Decision Tree Learning)

Η εκμάθηση διάρθρωσης αποφάσεων χρησιμοποιεί ένα δέντρο απόφασης (ως μοντέλο πρόβλεψης) για να μεταβεί από τις παρατηρήσεις για ένα στοιχείο (που αντιπροσωπεύεται στους κλάδους) σε συμπεράσματα σχετικά με την τιμή στόχου του αντικειμένου (που αναπαριστάται στα φύλλα). Πρόκειται για μία από τις προγνωστικές μεθόδους μοντελοποίησης που χρησιμοποιούνται στις στατιστικές, στην εξόρυξη δεδομένων και στη μηχανική μάθηση. Τα μοντέλα δένδρων όπου η μεταβλητή-στόχος μπορεί να πάρει ένα διακεκριμένο σύνολο τιμών ονομάζονται δέντρα ταξινόμησης. Σε αυτές τις δομές δέντρων, τα φύλλα αντιπροσωπεύουν ετικέτες τάξεων και κλάδους αντιπροσωπεύουν συζεύξεις χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσης. Τα δέντρα αποφάσεων όπου η μεταβλητή-στόχος μπορεί να πάρει συνεχείς τιμές (τυπικά πραγματικοί αριθμοί) ονομάζονται δέντρα παλινδρόμησης.

Στην ανάλυση αποφάσεων, ένα δέντρο απόφασης μπορεί να χρησιμοποιηθεί για την οπτική και ρητή εκπροσώπηση των αποφάσεων και της λήψης αποφάσεων. Στην εξόρυξη δεδομένων, ένα δέντρο απόφασης περιγράφει τα δεδομένα (αλλά το δέντρο ταξινόμησης που θα προκύψει μπορεί να αποτελέσει εισροή για τη λήψη αποφάσεων).

Η μάθηση δέντρων αποφάσεων είναι μια μέθοδος που χρησιμοποιείται συνήθως στην εξόρυξη δεδομένων. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που προβλέπει την τιμή μιας μεταβλητής στόχου με βάση διάφορες μεταβλητές εισόδου. Κάθε εσωτερικός κόμβος αντιστοιχεί σε μία από τις μεταβλητές εισόδου. Υπάρχουν άκρα στα παιδιά για κάθε μία από τις πιθανές τιμές αυτής της μεταβλητής εισόδου. Κάθε φύλλο αντιπροσωπεύει μια τιμή της μεταβλητής στόχου δεδομένων των τιμών των μεταβλητών εισόδου που αντιπροσωπεύει η διαδρομή από τη ρίζα στο φύλλο. Ένα δέντρο απόφασης είναι μια απλή αναπαράσταση για την ταξινόμηση παραδειγμάτων. Υποθέστε ότι όλες οι λειτουργίες εισόδου έχουν πεπερασμένους διακριτούς τομείς και υπάρχει ένα μόνο χαρακτηριστικό στόχου που ονομάζεται "ταξινόμηση". Κάθε στοιχείο του τομέα της ταξινόμησης ονομάζεται κλάση. Ένα δέντρο απόφασης ή ένα δέντρο ταξινόμησης είναι ένα δέντρο στο οποίο κάθε εσωτερικός (μη φύλλων) κόμβος φέρει ετικέτα με ένα χαρακτηριστικό εισόδου. Τα τόξα που προέρχονται από έναν κόμβο με ένα χαρακτηριστικό εισόδου επισημαίνονται με κάθε μία από τις πιθανές τιμές του χαρακτηριστικού στόχου ή εξόδου ή το τόξο οδηγεί σε έναν δευτερεύοντα κόμβο απόφασης σε διαφορετικό χαρακτηριστικό εισόδου. Κάθε φύλλο του δέντρου φέρει μια ταξινόμηση ή μια κατανομή πιθανότητας στις τάξεις. Ένα δέντρο μπορεί να «μάθει» διαιρώντας το σύνολο πηγών σε υποσύνολα με βάση μια δοκιμασία τιμής χαρακτηριστικού. Αυτή η διαδικασία επαναλαμβάνεται

σε κάθε παραγόμενο υποσύνολο με έναν αναδρομικό τρόπο που ονομάζεται αναδρομικός διαχωρισμός. Η επανάληψη ολοκληρώνεται όταν το υποσύνολο σε έναν κόμβο την ίδια τιμή της μεταβλητής στόχου ή όταν η διάσπαση δεν προσθέτει πλέον αξία στις προβλέψεις. Αυτή η διαδικασία καθοδήγησης των δέντρων αποφάσεων είναι ένα παράδειγμα αλγορίθμου και είναι μακράν η πιο κοινή στρατηγική για την εκμάθηση των δέντρων αποφάσεων από τα δεδομένα. Στην εξόρυξη δεδομένων, τα δέντρα απόφασης μπορούν επίσης να περιγραφούν ως συνδυασμός μαθηματικών και υπολογιστικών τεχνικών για την ενίσχυση της περιγραφής, κατηγοριοποίησης και γενίκευσης ενός δεδομένου συνόλου δεδομένων.

Η εκμάθηση δέντρων αποφάσεων είναι μία από τις πιο ευρέως χρησιμοποιούμενες και πρακτικές μεθόδους για την επαγωγική εξαγωγή συμπερασμάτων. Πρόκειται για μια μέθοδο προσέγγισης διακριτών λειτουργιών που είναι εύρωστη σε θορυβώδη δεδομένα και ικανή να μάθει διακριτικές εκφράσεις.

Η μάθηση δέντρων αποφάσεων είναι μια μέθοδος για την προσέγγιση των διακριτών λειτουργιών στόχου, στις οποίες η μαθησιακή συνάρτηση αντιπροσωπεύεται από ένα δέντρο απόφασης. Τα μαθημένα δέντρα μπορούν επίσης να αναπαρασταθούν εκ νέου ως σύνολο κανόνων if-then για να βελτιώσουν την αναγνωσιμότητα του ανθρώπου. Αυτές οι μέθοδοι μάθησης συγκαταλέγονται μεταξύ των πιο δημοφιλών επαγωγικών αλγορίθμων συμπερασμάτων και έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα καθηκόντων από τη μάθηση για τη διάγνωση ιατρικών περιπτώσεων έως τη μάθηση για την αξιολόγηση του πιστωτικού κινδύνου των αιτούντων δάνειο.

Τα δέντρα απόφασης ταξινομούν τα στιγμιότυπα διαχωρίζοντάς τα στο δέντρο από τη ρίζα σε κάποιο κόμβο φύλλων, το οποίο παρέχει την ταξινόμηση της στιγμής. Κάθε κόμβος στο δέντρο καθορίζει μια δοκιμή κάποιου χαρακτηριστικού της στιγμιότυπου και κάθε κλάδος που κατέρχεται από αυτόν τον κόμβο αντιστοιχεί σε μία από τις πιθανές τιμές αυτού του χαρακτηριστικού. Μια περίπτωση ταξινομείται ξεκινώντας από τον ριζικό κόμβο του δέντρου, δοκιμάζοντας το χαρακτηριστικό που καθορίζεται από αυτόν τον κόμβο και μετά μετακινώντας προς τα κάτω το κλαδί του δέντρου που αντιστοιχεί στην τιμή του χαρακτηριστικού στο συγκεκριμένο παράδειγμα. Αυτή η διαδικασία επαναλαμβάνεται στη συνέχεια στον νέο κόμβο[30]

Για την κατασκευή μερικές δέντρου αποφάσεων αναφέρεται η ακόλουθη αλγοριθμική διαδικασία:

1. Βήμα 1: Δημιουργία μερικές κόμβου που περιέχει μερικές μερικές εγγραφές.

2. Βήμα 2: Διάσπαση του κόμβου με βάση μια συνθήκη, έτσι ώστε να διαχωριστούν οι εγγραφές σε κάποιο από τα γνωρίσματα.

3. Βήμα 3: Γίνεται αναδρομική κλίση του βήματος 2 σε κάθε κόμβο μέχρι να φτάσουμε στον τελικό κόμβο, ώστε να υπάρξει η τελική απόφαση. Όταν κάποιος κόμβος δεν έχει παραδείγματα, τότε αντιστοιχίζεται σε μια κατηγορία. Αν σε κάποιο κόμβο, υπάρχουν θετικά και αρνητικά παραδείγματα, αλλά έχουν εξαντληθεί όλα τα χαρακτηριστικά, τότε ο κόμβος χαρακτηρίζεται διφορούμενος. Τότε ο κόμβος μερικές μπορεί να αντιστοιχηθεί στην πλειοψηφία κατηγορία παραδειγμάτων ή μερικές κατηγορίες με μερικές αντίστοιχες συχνότητες εμφάνισής του. Αυτό μπορεί να συμβεί είτε από την ύπαρξη θορύβου στα δεδομένα ή παράβλεψη σημαντικών χαρακτηριστικών.

4. Βήμα 4: Αφού κατασκευαστεί το δέντρο αποφάσεων, μπορούν να γίνουν κάποιες βελτιστοποιήσεις με τη μέθοδο κλαδέματος.

Κανόνες τερματισμού διαδικασίας:

Η διαδικασία διαχωρισμού των κόμβων στο δέντρο σταματάει με βάση κάποιους κανόνες. Όταν ικανοποιηθεί τουλάχιστον μερικές από μερικές παρακάτω κανόνες τερματισμού για κάθε κόμβο του δέντρου, τότε ο αλγόριθμος θα σταματήσει.

- Μερικές οι καταχωρήσεις έχουν την ίδια τιμή για το πεδίο – στόχο (ο κόμβος είναι καθαρός).

- Μερικές οι καταχωρήσεις στον κόμβο έχουν την ίδια τιμή για όλα τα πεδία πρόβλεψης τα οποία χρησιμοποιούνται από το μοντέλο.

- Το βάθος του δέντρου για τον τρέχον κόμβο είναι το μέγιστο βάθος που έχει προκαθοριστεί. Ο τρέχον κόμβος καθορίζεται από τον αριθμό των διαδοχικών διαχωρισμένων κόμβων.

- Ο αριθμός των καταχωρίσεων στον κόμβο είναι μικρότερος από το ελάχιστο μέγεθος του μητρικού κόμβου, μερικές αυτό έχει προκαθοριστεί.

- Ο αριθμός των καταχωρίσεων σε οποιονδήποτε από μερικές θυγατρικούς κόμβους που προκύπτουν από τον καλύτερο διαχωρισμό κόμβου, είναι μικρότερος από το ελάχιστο μέγεθος του θυγατρικού κόμβου, μερικές έχει προκαθοριστεί.

- Ο καλύτερος διαχωρισμός για τον κόμβο αποδίδει μια μείωση στην μη –καθαρότητα η οποία είναι μικρότερη από την ελάχιστη μεταβολή μη-καθαρότητας μερικές αυτή έχει προκαθοριστεί.

**6.1 Τα δέντρα αποφάσεων που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι δύο βασικών τύπων:**

- **Δέντρα Ταξινόμησης**

Η ανάλυση δένδρων ταξινόμησης (classification tree) είναι όταν το προβλεπόμενο αποτέλεσμα είναι η κλάση στην οποία ανήκουν τα δεδομένα.

- **Δέντρα Παλινδρόμησης**

Η ανάλυση δέντρων παλινδρόμησης (regression tree) είναι όταν το προβλεπόμενο αποτέλεσμα μπορεί να θεωρηθεί πραγματικός αριθμός (π.χ. η τιμή ενός σπιτιού ή η διάρκεια παραμονής ενός ασθενούς σε ένα νοσοκομείο)

### 6.2 Πλεονεκτήματα και Μειονεκτήματα των Δέντρων Αποφάσεων

**Πλεονεκτήματα:**

- Τα δέντρα αποφάσεων παράγουν μοντέλα που ερμηνεύονται εύκολα.

- Είναι μη-παραμετρικά και κατά συνέπεια δεν προϋποθέτουν τα δεδομένα ν' ακολουθούν την κανονική κατανομή.

- Μπορούν να χειριστούν όλων των ειδών τα δεδομένα (συνεχή, κατηγορικά, δυαδικά, διατακτικά), χωρίς να απαιτούν κάποιον μετασχηματισμό.

- Δεν απαιτούν την προεπιλογή των μεταβλητών, διότι ο αλγόριθμος ξεχωρίζει τις πιο σημαντικές αγνοώντας τις υπόλοιπες.

**Μειονέκτημα** είναι η ευαισθησία που παρουσιάζουν τα δέντρα ταξινόμησης σε μικρές μεταβολές στα δεδομένα εκπαίδευσης.

Όταν παραλείπονται μερικές παρατηρήσεις, υπάρχει η πιθανότητα να αλλάξουν οι μεταβλητές διαχωρισμού και να προκύψουν διαφορετικά δέντρα[41].

Σε θέματα ιατρικά, όπως στη συγκεκριμένη διπλωματική εργασία, είναι προφανές ότι η εσφαλμένα αρνητική ταξινόμηση έχει χειρότερες συνέπειες από ότι η εσφαλμένη θετική.

### 6.3 Δέντρο ταξινόμησης για την αξιολόγηση του ρόλου των δημογραφικών δεδομένων και των συμπτωμάτων στον κίνδυνο εμφάνισης καρκίνου στις κύστεις των ωοθηκών

Τα δέντρα ταξινόμησης είναι μία από τις πιο δημοφιλείς ταξινομήσεις στους αλγορίθμους μάθησης. Η μέθοδος αυτή παίρνει είσοδο ένα διάνυσμα τιμών σε κάποιες ιδιότητες

και επιστρέφει μία έξοδο. Αυτή η έξοδος μπορεί να είναι διακριτή, οπότε ορίζεται ένα πρόβλημα ταξινόμησης, ενώ αν η έξοδος είναι συνεχής υπάρχει πρόβλημα παλινδρόμησης.

Σκοπός της διπλωματικής εργασίας και μελέτης ήταν να αξιολογηθεί, με χρήση δέντρου ταξινόμησης, η δομή σχέσεων μεταξύ δημογραφικών δεδομένων και συμπτωμάτων από γυναίκες με την εμφάνιση καρκίνου στις κύστες των ωοθηκών.

Τα δέντρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων χαρακτηριστικών. Γενικά τα δέντρα απόφασης έχουν καλή ακρίβεια.

Ένα δέντρο απόφασης αναπαριστά μια διαδικασία λήψης απόφασης, όπου για κάθε πιθανό σημείο ή κατάσταση έχουμε ένα κόμβο, ενώ για κάθε πιθανή επιλογή που μπορεί να γίνει σε ένα σημείο απόφασης αναπαρίσταται με ένα «κόμβο-παιδί». Κάθε κόμβος ορίζει μία συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων. Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο. Στα κλαδιά καταλήγουν οι τελικοί κόμβοι που ανήκουν σε ένα μόνο σύνολο, όπου είναι οι τελικές αποφάσεις ή ενέργειες[31].

Τα δέντρα ταξινόμησης λειτουργούν ακριβώς σαν δέντρα παλινδρόμησης, μόνο που προσπαθούν να προβλέψουν μια διακριτή κατηγορία (την κλάση) και όχι μια αριθμητική τιμή.

Οι μεταβλητές που εμπίπτουν στην ταξινόμηση μπορούν να είναι αριθμητικές ή κατηγορηματικές, το ίδιο ισχύει και σε ένα δέντρο παλινδρόμησης. Παρέχουν αρκετά κατανοητούς προγνωστικούς παράγοντες σε καταστάσεις όπου υπάρχουν πολλές μεταβλητές που αλληλεπιδρούν με περίπλοκο, μη γραμμικό τρόπο. Βρήκαμε τα δέντρα ταξινόμησης σχεδόν με τον ίδιο τρόπο που βρήκαμε τα δέντρα παλινδρόμησης: ξεκινάμε με έναν μόνο κόμβο και στη συνέχεια ψάχνουμε για τη δυαδική διάκριση που μας δίνει τις περισσότερες πληροφορίες για την τάξη. Στη συνέχεια, λαμβάνουμε κάθε έναν από τους νέους κόμβους που προκύπτουν και επαναλαμβάνουμε τη διαδικασία εκεί, συνεχίζοντας την επανάληψη μέχρι να φτάσουμε σε κάποιο κριτήριο διακοπής. Το δέντρο που προκύπτει συχνά είναι υπερβολικά μεγάλο, γι’ αυτό το κλαδεύουμε ξανά χρησιμοποιώντας (για παράδειγμα) την εγκάρσια επικύρωση.

*Οι διαφορές από την καλλιέργεια παλινδρόμησης έχουν να κάνουν με:*

- Τον τρόπο μέτρησης των πληροφοριών.

- Τι είδους προβλέψεις κάνει το δέντρο.
- Πώς μετράμε το προγνωστικό σφάλμα[47].

### 6.4 Random Forest

Τα τυχαία δάση (Random Forest) είναι μία από τις πιο αποτελεσματικές μεθόδους. Μέθοδος ονομασμένη από τους Leo Breiman και Adele Cutler. Η μέθοδος αυτή ανήκει στην κατηγορία των μεθόδων που χρησιμοποιούν μια κλασική μέθοδο κατηγοριοποίησης πολλές φορές προκειμένου να ενισχύσουν την αποτελεσματικότητα της συνολικής μεθόδου[40]. Είναι εύκολα κατανοητό ότι η ονομασία «δάση» προέρχεται από το γεγονός της κατασκευής πολλών δέντρων απόφασης. Στη συνέχεια θα δούμε γιατί αποκαλούνται και «τυχαία».

Οι ασθενείς(weak), όπως αποκαλούνται, οι επιμέρους κατηγοριοποιητές μπορεί να είναι οποιοδήποτε εργαλείο ή μέθοδος κατηγοριοποίησης, όπως π.χ. στην περίπτωσή μας είναι δέντρα απόφασης. Τα τυχαία δάση, όπως όλες οι μέθοδοι που χρησιμοποιούν κάποιο τύπο boosting, λειτουργούν με παρόμοιο τρόπο. Αφού κατασκευαστούν με κάποιο τρόπο οι επιμέρους ασθενείς κατηγοριοποιητές, ο κάθε ένας ψηφίζει για την κλάση που ανήκει η παρατήρηση προς κατηγοριοποίηση. Η κλάση που επικρατεί είναι αυτή που παίρνει τους περισσότερους ψήφους. Ενώ ο κάθε κατηγοριοποιητής από μόνος του έχει ανακριβή συνήθως αποτελέσματα και χαμηλή αξιοπιστία, σαν σύνολο πετυχαίνουν από τις πιο ακριβείς κατηγοριοποιήσεις ανάμεσα σε όλες τις μεθόδους.

Το κάθε δέντρο που κατασκευάζεται δεν μοιάζει τόσο πολύ σε ένα προσεγμένο δέντρο απόφασης. Στην περίπτωση των τυχαίων δασών, αρχικά επιλέγεται ένα τυχαίο μικρό υπόδειγμα όλων των χαρακτηριστικών της βάσης και στη συνέχεια χρησιμοποιούνται τυχαία σε κάθε κόμβο μέχρι να κατασκευαστεί το δέντρο. Ύστερα επιλέγεται ένα τυχαίο δείγμα παρατηρήσεων από την βάση για training, και το υπόλοιπο χρησιμοποιείται για να εκτιμήσει το σφάλμα του δέντρου. Η μέθοδος επαναλαμβάνεται για τον αριθμό των δέντρων που έχουμε καθορίσει. Να σημειωθεί επίσης ότι τα δέντρα δεν υποβάλλονται σε αποκοπή κόμβων και μεγαλώνουν όσο είναι δυνατόν.

*Οι δημιουργοί των τυχαίων δασών αναφέρουν ότι το πόσο καλός είναι ένας κατηγοριοποιητής τυχαίων δασών στηρίζεται σε δύο πράγματα:*

- Στην επιρροή του κάθε δέντρου. Η ικανότητα του κάθε δέντρου έχει να κάνει με το συνολικό σφάλμα του, δηλαδή ένα δέντρο με χαμηλό σφάλμα έχει περισσότερη δύναμη και επηρεάζει θετικά τον ρυθμό σφάλματος του δάσους.
- Στην συσχέτιση που μπορεί να έχουν δύο δέντρα μεταξύ τους. Η συσχέτιση έχει να κάνει με το πόσο μοιάζουν δύο δέντρα, δηλαδή αν χρησιμοποιούν κοινά χαρακτηριστικά στην κατασκευή τους και είναι ανάλογη του ρυθμού σφάλματος.

Ο αριθμός χαρακτηριστικών που θα επιλεγούν για κάθε δέντρο επηρεάζει τους παραπάνω παράγοντες. Λιγότερα χαρακτηριστικά σημαίνει λιγότερο συσχετισμένα δέντρα, αλλά μεγαλύτερο συνολικό σφάλμα και αντιστρόφως. Μπορεί να βρεθεί ένα πεδίο για τον αριθμό αυτό, το οποίο αν τηρείται ο αλγόριθμος έχει μικρότερο δυνατό σφάλμα.

Ενώ τα τυχαία δάση έχουν πολλά προτερήματα, όπως υψηλή ακρίβεια, καλή αξιοπιστία σε μεγάλες βάσεις και θεώρηση όλων των χαρακτηριστικών τους, φαίνεται ότι είναι ευάλωτα σε overfitting σε συγκεκριμένες βάσεις, αντίθετα με το τι δηλώνουν οι δημιουργοί της μεθόδου. Επίσης, ένα άλλο αρνητικό σημείο τους είναι το γεγονός ότι δεν μπορούν να δημιουργηθούν ίδια δάση, εξαιτίας του τυχαίου τρόπου με τον οποίο δημιουργούνται, που σημαίνει ότι «ικανότερα» δέντρα δεν μπορούν να επαναληφθούν για καλύτερη ακρίβεια[42].

Πιο αναλυτικά κάθε δέντρο εξαρτάται από τις τιμές ενός δειγματοληπτικού φορέα δειγματοληψίας ανεξάρτητα και με την ίδια κατανομή για όλα τα δέντρα στο δάσος. Κατά την ταξινόμηση, κάθε δέντρο ψηφίζει και η πιο δημοφιλής τάξη επιστρέφεται. Τα τυχαία δάση μπορούν να κατασκευαστούν με τη χρήση σάκων σε συνδυασμό με την επιλογή τυχαίων χαρακτηριστικών. Παρέχεται ένα σετ κατάρτισης, D, των d πλειάδων.

*Η γενική διαδικασία για τη δημιουργία δέντρων απόφασης k για το σύνολο είναι η εξής:*

Για κάθε επανάληψη, i (i = 1, 2, ..., k) λαμβάνεται δειγματοληπτικό σύνολο Di, των d tuples με αντικατάσταση από το D. Δηλαδή κάθε Di είναι ένα δείγμα bootstrap του D, έτσι ώστε ορισμένες πλειάδες να μπορούν να εμφανιστούν περισσότερες από μία φορές στο Di, ενώ άλλες

μπορεί να αποκλειστούν. Έστω F ο αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν για τον προσδιορισμό του διαχωρισμού σε κάθε κόμβο, όπου το F είναι πολύ μικρότερο από τον αριθμό των διαθέσιμων ιδιοτήτων. Για να κατασκευάσουμε έναν ταξινομητή δέντρων αποφάσεων, Mi, επιλέγουμε τυχαία σε κάθε κόμβο τις ιδιότητες F ως υποψήφιους για τη διάσπαση στον κόμβο. Η μεθοδολογία CART χρησιμοποιείται για την ανάπτυξη των δέντρων. Τα δέντρα μεγαλώνουν στο μέγιστο μέγεθος και δεν κλαδεύονται. Τυχαία δάση που σχηματίζονται με αυτόν τον τρόπο, με τυχαία επιλογή εισόδου, ονομάζονται Forest-RI. Μια άλλη μορφή τυχαίου δάσους, που ονομάζεται Forest-RC, χρησιμοποιεί τυχαίους γραμμικούς συνδυασμούς των χαρακτηριστικών εισόδου. Αντί να επιλέγει τυχαία ένα υποσύνολο των χαρακτηριστικών, δημιουργεί νέα χαρακτηριστικά (ή χαρακτηριστικά) που είναι ένας γραμμικός συνδυασμός των υπαρχόντων χαρακτηριστικών. Δηλαδή, ένα χαρακτηριστικό παράγεται με τον προσδιορισμό του L, τον αριθμό των πρωτότυπων χαρακτηριστικών που θα συνδυαστούν. Σε έναν δεδομένο κόμβο, τα χαρακτηριστικά L επιλέγονται τυχαία και προστίθενται μαζί με συντελεστές που είναι ομοιόμορφοι τυχαίοι αριθμοί στο [-1,1]. Οι γραμμικοί συνδυασμοί F δημιουργούνται και γίνεται αναζήτηση πάνω σε αυτές για τον καλύτερο διαχωρισμό. Αυτή η μορφή τυχαίου δάσους είναι χρήσιμη όταν υπάρχουν μόνο λίγα χαρακτηριστικά διαθέσιμα, έτσι ώστε να μειωθεί η συσχέτιση μεταξύ μεμονωμένων ταξινομητών.

Τα τυχαία δάση είναι συγκρίσιμα με την ακρίβεια με το AdaBoost, αλλά είναι πιο ανθεκτικά στα σφάλματα. Το σφάλμα γενίκευσης για ένα δάσος συγκλίνει, εφόσον ο αριθμός των δένδρων στο δάσος είναι μεγάλος. Έτσι, η υπερφόρτωση δεν είναι πρόβλημα. Η ακρίβεια ενός τυχαίου δάσους εξαρτάται από τη δύναμη των μεμονωμένων ταξινομητών και από ένα μέτρο της εξάρτησης μεταξύ τους. Το ιδανικό είναι να διατηρηθεί η αντοχή των μεμονωμένων ταξινομητών χωρίς να αυξηθεί ο συσχετισμός τους. Τα τυχαία δάση δεν είναι ευαίσθητα στον αριθμό των χαρακτηριστικών που επιλέγονται για εξέταση σε κάθε διάσπαση. Συνήθως, επιλέγονται μέχρι το log2d + 1. Επειδή τα τυχαία δάση θεωρούν πολύ λιγότερα χαρακτηριστικά για κάθε διαίρεση, είναι αποτελεσματικά σε πολύ μεγάλες βάσεις δεδομένων. Τα τυχαία δάση παρέχουν εσωτερικές εκτιμήσεις μεταβλητής σημασίας[31].

**Χαρακτηριστικά Τυχαίων Δασών:**

- Είναι εξαιρετικά στην ακρίβεια μεταξύ των τρεχόντων αλγορίθμων.
- Λειτουργούν αποτελεσματικά σε μεγάλες βάσεις δεδομένων.
- Μπορούν να χειριστούν χιλιάδες μεταβλητές εισόδου χωρίς τη διαγραφή μεταβλητής.
- Δίνουν εκτιμήσεις για το ποιες μεταβλητές είναι σημαντικές στην ταξινόμηση.
- Δημιουργούν μια εσωτερική αμερόληπτη εκτίμηση του σφάλματος γενίκευσης καθώς το κτίριο των δασών εξελίσσεται.
- Έχουν μια αποτελεσματική μέθοδο για την εκτίμηση των ελλειπόντων δεδομένων και διατηρεί την ακρίβεια όταν λείπει μεγάλο ποσοστό των δεδομένων.
- Έχουν μεθόδους για την εξισορρόπηση σφαλμάτων στα σύνολα δεδομένων που δεν έχουν ισοσταθμιστεί.
- Τα παραγόμενα δάση μπορούν να αποθηκευτούν για μελλοντική χρήση σε άλλα δεδομένα.
- Χρησιμοποιούνται πρωτότυπα που δίνουν πληροφορίες σχετικά με τη σχέση μεταξύ των μεταβλητών και της ταξινόμησης.
- Υπολογίζουν τις γειτνιάζουσες μεταξύ ζευγών περιπτώσεων που μπορούν να χρησιμοποιηθούν στη συσσωμάτωση, εντοπίζοντας ακραίες τιμές, ή (με κλιμάκωση) δίνουν ενδιαφέρουσες απόψεις των δεδομένων.
- Οι δυνατότητες των παραπάνω μπορούν να επεκταθούν σε μη επισημασμένα δεδομένα, οδηγώντας σε μη ομαδοποιημένη ομαδοποίηση, σε προβολές δεδομένων και ανίχνευση εξωστρέφειας.
- Προσφέρουν μια πειραματική μέθοδο ανίχνευσης μεταβλητών αλληλεπιδράσεων.

Τα τυχαία δάση είναι γρήγορα και μπορείτε να τρέξετε όσο περισσότερα δέντρα θέλετε. Για παράδειγμα, λειτουργώντας σε ένα σύνολο δεδομένων με 50.000 περιπτώσεις και 100 μεταβλητές, μπορούν να παράγουν 100 δέντρα σε 11 λεπτά σε μια μηχανή 800Mhz. Για τα μεγάλα σύνολα δεδομένων, η μεγάλη απαίτηση μνήμης είναι η αποθήκευση των ίδιων των δεδομένων και τριών ακέραιων συστοιχιών με τις ίδιες διαστάσεις με τα δεδομένα. Εάν

υπολογιστούν οι εγγύτητες, οι απαιτήσεις αποθήκευσης αυξάνονται ανάλογα με τον αριθμό των περιπτώσεων με τον αριθμό των δέντρων.

**Αλληλεπιδράσεις**

Ο λειτουργικός ορισμός της αλληλεπίδρασης που χρησιμοποιείται είναι ότι οι μεταβλητές m και k αλληλεπιδρούν σε μία μεταβλητή, για παράδειγμα, η μεταβλητή m σε ένα δέντρο κάνει ένα διαχωρισμό στο k συστηματικά λιγότερο πιθανό ή πιο πιθανό. Η εφαρμογή που χρησιμοποιείται βασίζεται στις τιμές gini g (m) για κάθε δέντρο στο δάσος. Αυτά ταξινομούνται για κάθε δέντρο και για κάθε δύο μεταβλητές, η απόλυτη διαφορά των τάξεων τους υπολογίζεται κατά μέσο όρο σε όλα τα δέντρα.

Αυτός ο αριθμός υπολογίζεται επίσης υπό την υπόθεση ότι οι δύο μεταβλητές είναι ανεξάρτητες η μία από την άλλη και οι τελευταίες αφαιρούνται από την πρώτη. Ένας μεγάλος θετικός αριθμός υποδηλώνει ότι μια διάσπαση σε μια μεταβλητή εμποδίζει μια διάσπαση από την άλλη και αντιστρόφως. Πρόκειται για μια πειραματική διαδικασία της οποίας τα συμπεράσματα πρέπει να εξεταστούν με προσοχή. Έχει δοκιμαστεί μόνο σε λίγα σύνολα δεδομένων.

**Μεγάλη σημασία (Gini Importance)**

Κάθε φορά που γίνεται διάσπαση ενός κόμβου σε μεταβλητή m, το κριτήριο προσμίξεων gini για τους δύο απογόνους είναι μικρότερο από τον γονικό κόμβο. Η προσθήκη των μειώσεων του gini για κάθε μεμονωμένη μεταβλητή σε όλα τα δέντρα του δάσους δίνει μια γρήγορη μεταβλητή σημασία που είναι συχνά πολύ συνεπής με το μέτρο σημασίας της μετάθεσης.

**Εγγύτητα**

Αυτά είναι ένα από τα πιο χρήσιμα εργαλεία σε τυχαία δάση. Σχηματίστηκε αρχικά μια μήτρα NxN. Αφού καλλιεργηθεί ένα δέντρο, βάλτε όλα τα δεδομένα, τόσο την κατάρτιση όσο και το δέντρο, κάτω από το δέντρο. Εάν οι περιπτώσεις k και n βρίσκονται στον ίδιο τερματικό κόμβο, αυξάνεται η εγγύτητα τους κατά ένα. Στο τέλος, ομαλοποιήστε τις γειτνιάσεις διαιρώντας τον αριθμό των δέντρων.

Οι χρήστες σημείωσαν ότι με μεγάλα σύνολα δεδομένων δεν μπορούσαν να χωρέσουν ένα matrix NxN σε γρήγορη μνήμη. Μία τροποποίηση μείωσε το απαιτούμενο μέγεθος μνήμης

σε NxT όπου T είναι ο αριθμός των δένδρων στο δάσος. Για να επιταχυνθεί η αντικατάσταση της κλίμακας έντασης υπολογισμού και επαναληπτικής έλλειψης αξίας, ο χρήστης έχει τη δυνατότητα να διατηρεί μόνο τις μεγαλύτερες οριακές τιμές σε κάθε περίπτωση.

Όταν υπάρχει ένα σετ δοκιμών, μπορούν επίσης να υπολογιστούν οι γειτνιάσεις κάθε θήκης στο σετ δοκιμών με κάθε περίπτωση στο σετ εκπαίδευσης. Το ποσό των πρόσθετων υπολογιστών είναι μέτριο.

Κάθε δέντρο καλλιεργείται ως εξής:

1. Αν ο αριθμός των περιπτώσεων στο σετ εκπαίδευσης είναι N, δείγματα N τυχαία - αλλά με αντικατάσταση, από τα αρχικά δεδομένα, αυτό το δείγμα θα είναι το σετ κατάρτισης για την καλλιέργεια του δέντρου.

2. Εάν υπάρχουν μεταβλητές εισόδου M, ένας αριθμός m<<M καθορίζεται έτσι ώστε σε κάθε κόμβο, οι μεταβλητές m επιλέγονται τυχαία από το M και ο καλύτερος διαχωρισμός σε αυτά τα m χρησιμοποιείται για τον διαχωρισμό του κόμβου. Η τιμή του m διατηρείται σταθερή κατά τη διάρκεια της δασικής ανάπτυξης.

3. Κάθε δέντρο καλλιεργείται στον μεγαλύτερο δυνατό βαθμό. Δεν υπάρχει κλάδεμα.

Δύο πολύ ωραίες ιδιότητες των τυχαίων δασών είναι:

• Μπορούν να χρησιμοποιηθούν τα δεδομένα για να πάρουμε μια αμερόληπτη εκτίμηση του σφάλματος ταξινόμησης.

• Είναι εύκολο να υπολογιστεί ένα μέτρο με "μεταβλητή σημασία".

*Το ποσοστό σφάλματος εξαρτάται από δύο πράγματα:*

1. Την συσχέτιση μεταξύ οποιωνδήποτε δύο δένδρων στο δάσος. Η αύξηση της συσχέτισης αυξάνει το ποσοστό σφάλματος.
2. Η αντοχή κάθε δέντρου στο δάσος. Ένα δέντρο με χαμηλό ποσοστό σφάλματος είναι ένας ισχυρός ταξινομητής. Η αύξηση της αντοχής των μεμονωμένων δένδρων μειώνει το ποσοστό σφάλματος. Η μείωση m μειώνει τόσο τη συσχέτιση όσο και τη δύναμη. Η αύξηση αυξάνεται και στα δύο. Κάπου στο μέσον είναι μια "βέλτιστη" περιοχή m - συνήθως αρκετά μεγάλη. Αυτή είναι η μόνη ρυθμιζόμενη παράμετρος στην οποία τα τυχαία δάση είναι κάπως ευαίσθητα[48].

# 7. Στατιστική Μάθηση (Statistical Learning)

## 7.1 Εισαγωγή

Στις αρχές του δέκατου ένατου αιώνα, οι Legendre και Gauss δημοσίευσαν έγγραφα σχετικά με τη μέθοδο των ελαχίστων τετραγώνων, τα οποία εφάρμοζαν την παλαιότερη μορφή του γνωστού ως γραμμική παλινδρόμηση. Η προσέγγιση εφαρμόστηκε πρώτα με επιτυχία σε προβλήματα αστρονομίας. Η γραμμική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη ποσοτικών τιμών. Προκειμένου να προβλεφθούν ποιοτικές τιμές, όπως εάν ο ασθενής επιβιώνει ή πεθαίνει, η Fisher πρότεινε γραμμική ανάλυση διακρίσεων το 1936. Στη δεκαετία του 1940, διάφοροι συγγραφείς παρουσίασαν μια εναλλακτική προσέγγιση, την υλικοτεχνική παλινδρόμηση. Στις αρχές της δεκαετίας του '70, οι Nelder και Wedderburn επινόησαν τον όρο γενικευμένα γραμμικά μοντέλα για μια ολόκληρη τάξη μεθόδων στατιστικής μάθησης που περιλαμβάνουν τόσο γραμμική όσο και λογική παλινδρόμηση ως ειδικές περιπτώσεις.

Μέχρι τα τέλη της δεκαετίας του 1970, ήταν διαθέσιμες πολλές άλλες τεχνικές μάθησης από δεδομένα. Εντούτοις, ήταν σχεδόν αποκλειστικά γραμμικές μέθοδοι, επειδή οι μη γραμμικές σχέσεις προσαρμογής δεν μπορούσαν να υπολογιστούν τότε. Μέχρι τη δεκαετία του 1980, η τεχνολογία πληροφορικής είχε τελικά βελτιωθεί επαρκώς ώστε οι μη γραμμικές μέθοδοι δεν ήταν πλέον απαγορευτικές από υπολογιστική άποψη. Στα μέσα της δεκαετίας του 1980, οι Breiman, Friedman, Olshen και Stone εισήγαγαν δέντρα ταξινόμησης και παλινδρόμησης και ήταν μεταξύ των πρώτων που επέδειξαν τη δύναμη μιας λεπτομερούς πρακτικής εφαρμογής μιας μεθόδου, συμπεριλαμβανομένης της εγκάρσιας επικύρωσης για την επιλογή μοντέλου. Οι Hastie και Tibshirani εξειδίκευαν τον όρο γενικευμένα πρόσθετα μοντέλα το 1986 για μια κατηγορία μη γραμμικών επεκτάσεων σε γενικευμένα γραμμικά μοντέλα και επίσης παρείχαν μια πρακτική εφαρμογή λογισμικού.

Από τότε, εμπνευσμένο από την έλευση της μηχανικής μάθησης και άλλων κλάδων, η στατιστική μάθηση έχει αναδειχθεί ως ένα νέο υπό πεδίο στα στατιστικά στοιχεία, επικεντρωμένο στην επιτηρούμενη και ανεξέλεγκτη μοντελοποίηση και πρόβλεψη. Τα τελευταία

χρόνια, η πρόοδος στη στατιστική μάθηση χαρακτηρίστηκε από την αυξανόμενη διαθεσιμότητα ισχυρού και σχετικά φιλικού προς το χρήστη λογισμικού, όπως το δημοφιλές και ελεύθερα διαθέσιμο σύστημα R. Αυτό έχει τη δυνατότητα να συνεχίσει τη μετατροπή του πεδίου από ένα σύνολο τεχνικών που χρησιμοποιούνται και αναπτύσσονται από τους στατιστικολόγους και τους επιστήμονες υπολογιστών σε ένα ουσιαστικό εργαλείο εργαλείων για μια πολύ ευρύτερη κοινότητα.

Η θεωρία της στατιστικής μάθησης είναι ένα πλαίσιο για την εκμάθηση μηχανικής μάθησης από τα πεδία των στατιστικών και της λειτουργικής ανάλυσης. Ασχολείται με το πρόβλημα της εύρεσης μιας προγνωστικής λειτουργίας με βάση τα δεδομένα. Η θεωρία της στατιστικής μάθησης έχει οδηγήσει σε επιτυχείς εφαρμογές σε τομείς όπως η όραση στον υπολογιστή, η αναγνώριση ομιλίας και η βιοπληροφορική.

Η στατιστική μάθηση αναφέρεται σε ένα τεράστιο σύνολο εργαλείων για την πρόβλεψη και κατανόηση των δεδομένων. Αυτά τα εργαλεία μπορούν να ταξινομηθούν ως εποπτευόμενα ή μη εποπτευόμενα. Από την άποψη της θεωρίας της στατιστικής μάθησης, η επίβλεψη της μάθησης είναι κατανοητή. Η εποπτευόμενη μάθηση περιλαμβάνει τη μάθηση από ένα σύνολο εκπαιδευτικών δεδομένων. Κάθε σημείο της εκπαίδευσης είναι ένα ζεύγος εισόδου-εξόδου, όπου η είσοδος χαρτώνεται σε μια έξοδο. Το μαθησιακό πρόβλημα συνίσταται στην εξαγωγή της συνάρτησης που χαρτογραφεί μεταξύ της εισόδου και της εξόδου, έτσι ώστε η μαθησιακή συνάρτηση να μπορεί να χρησιμοποιηθεί για την πρόβλεψη εξόδου από μελλοντική είσοδο.

Ανάλογα με τον τύπο της παραγωγής, τα προβλήματα εποπτευόμενης μάθησης είναι είτε προβλήματα παλινδρόμησης είτε προβλήματα ταξινόμησης[41].

Η στατιστική είναι μία μεθοδική μαθηματική, παλαιότερα τεχνική και σήμερα επιστήμη που επιχειρεί να εξαγάγει έγκυρη γνώση χρησιμοποιώντας εμπειρικά δεδομένα παρατήρησης ή και πειράματος. Κύριο αντικείμενο έρευνας και μελέτης της στατιστικής είναι η **συλλογή**, **ταξινόμηση**, **επεξεργασία**, **παρουσίαση**, **ανάλυση** και **ερμηνεία** διαφόρων δεδομένων με απώτερο στόχο την εξαγωγή ασφαλών συμπερασμάτων για λήψη ορθών αποφάσεων. Ως ιδιαίτερος κλάδος των μαθηματικών στην ουσία προσφέρει δύο σπουδαίες δυνατότητες αφενός την περιγραφή αριθμητικών συνόλων δεδομένων έρευνας και στη συνέχεια την ανάλυση αυτών. Συνέπεια αυτών των δυνατοτήτων είναι και η βασική διάκρισή της σε περιγραφική στατιστική και σε αναλυτική στατιστική.

- Στη **Περιγραφική στατιστική** περιγράφονται τα διάφορα στατιστικά στοιχεία μετά από συλλογή και ταξινόμηση κατά ομάδες των στατιστικών δεδομένων τα οποία ακολούθως παρουσιάζονται υπό μορφή ανάλυσης σε πίνακες, διαγράμματα με χαρακτηριστικές τιμές, ή ιδιότητες.
- Στην **Αναλυτική στατιστική**, που είναι περισσότερο περίπλοκη, αναζητείται με διάφορες μεθόδους ο προσδιορισμός βαθμού εμπιστοσύνης στην εξαγωγή ασφαλών συμπερασμάτων μέσα όμως από κάποιο περιορισμένο δείγμα στοιχείων ενός γενικότερου συνόλου.

**Τα βασικότερα στοιχεία της στατιστικής ανάλυσης είναι:**

- **Ο Πληθυσμός** (Population): είναι ένα σύνολο στοιχείων που μας ενδιαφέρει να μελετήσουμε ως προς ένα ή περισσότερα χαρακτηριστικά του.
- **Οι Μεταβλητές** ονομάζονται τα χαρακτηριστικά́ εκείνα, ως προς τα οποία εξετάζουμε έναν πληθυσμό και χωρίζονται σε δύο κατηγορίες, στις ποιοτικές και στις ποσοτικές.

    - **Ποιοτικές ή Ονομαστικές (nominal) μεταβλητές** είναι εκείνες που δεν επιδέχονται μέτρηση και οι τιμές τους δεν είναι αριθμοί́.
    - **Ποσοτικές ή Αριθμητικές μεταβλητές** είναι εκείνες που επιδέχονται μέτρηση και οι τιμές τους είναι αριθμοί́.

    Οι ποσοτικές μεταβλητές διακρίνονται σε συνέχεις και διακριτές (ασυνεχείς).

    - **Συνεχείς (scale)** είναι οι ποσοτικές μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή́ ενός διαστήματος ( α , β ).
    - **Διακριτές (ordinal)** είναι οι ποσοτικές μεταβλητές που παίρνουν μόνο μεμονωμένες τιμές.

    Οι ποιοτικές μεταβλητές διακρίνονται σε διατάξιμες και κατηγορικές.

    - **Διατάξιμες** είναι ποιοτικές μεταβλητές, οι οποίες λαμβάνουν ως τιμές έννοιες που ιεραρχούνται – ιεραρχημένες κατηγορίες.

- **Κατηγορικές** είναι ποιοτικές μεταβλητές, οι οποίες λαμβάνουν τιμές - έννοιες που δεν ιεραρχούνται - κατηγορίες χωρίς ταξινόμηση.

- **Οι τιμές της μεταβλητής**: είναι οι δυνατές τιμές που μπορεί να πάρει μία μεταβλητή.
- **Το Δείγμα** (Sample): είναι ένα υποσύνολο ενός πληθυσμού ή παρατήρηση αποτελεσμάτων μιας διαδικασίας για μια χρονική περίοδο.
- **Η Παράμετρος**(Parameter): είναι μία αριθμητική ποσότητα που συνοψίζει κάποιο χαρακτηριστικό του πληθυσμού ή της ικανότητας μιας διαδικασίας.
- **Η Στατιστική συνάρτηση**(Statistic): είναι μία αριθμητική ποσότητα που συνοψίζει κάποιο χαρακτηριστικό του δείγματος και που μπορεί να χρησιμοποιηθεί για την εκτίμηση μιας άγνωστης αντίστοιχης παραμέτρου του πληθυσμού.
- **Η Εμπιστοσύνη**(Confidence): είναι η πιθανοφάνεια ότι η στατιστική συμπερισματολογία στην οποία καταλήξαμε είναι σωστή ή ότι έχει κάποιο λάθος το οποίο όμως δεν υπερβαίνει κάποια προκαθορισμένη ποσότητα.
- **Το Διάστημα εμπιστοσύνης**(Confidence Interval): είναι η εκτίμηση ενός διαστήματος πιθανών τιμών με τη χρήση του δείγματος για μία άγνωστη παράμετρο του πληθυσμού.

### 7.2 Στατιστική Ανάλυση Κατηγορικών Μεταβλητών

Στην παρούσα διπλωματική εργασία θα μας απασχολήσουν τα **ποιοτικά δεδομένα.**

Οι μετρήσεις ποιοτικών δεδομένων γίνονται με τη βοήθεια κατηγοριών, όπου σε κάθε κατηγορία δηλώνετε ένα χαρακτηριστικό, το οποίο δεν μπορεί να αποτυπωθεί με αριθμό, όπως για παράδειγμα, η οικογενειακή κατάσταση, το φύλο, το είδος εργασίας. Οι τυχαίες μεταβλητές οι οποίες αποτελούνται από ποιοτικά δεδομένα λέγονται **κατηγορικές μεταβλητές** (categorical variable).

Μια κατηγορική μεταβλητή έχει μια κλίμακα μέτρησης που αποτελείται από ένα σύνολο κατηγοριών. Οι διαγνώσεις σχετικά με τον καρκίνο των ωοθηκών που βασίζονται σε υπέρηχο χρησιμοποιούν τις κατηγορίες, κανονικές, καλοήθεις, πιθανώς καλοήθεις, ύποπτες και κακοήθεις.

Η ανάπτυξη μεθόδων για κατηγορηματικές μεταβλητές διεγείρεται από ερευνητικές μελέτες στις κοινωνικές και βιοϊατρικές επιστήμες. Οι κατηγορικές κλίμακες είναι διαδεδομένες στις κοινωνικές επιστήμες για τη μέτρηση των στάσεων και των απόψεων. Οι καταγραφικές κλίμακες στις βιοϊατρικές επιστήμες μετρούν τα αποτελέσματα όπως είναι το εάν μια ιατρική θεραπεία είναι επιτυχής.

Αν και τα κατηγορικά δεδομένα είναι κοινά στις κοινωνικές και βιοϊατρικές επιστήμες, δεν περιορίζονται σε αυτές τις περιοχές. Συχνά εμφανίζονται στις επιστήμες συμπεριφοράς, τον τύπο ψυχικής ασθένειας, με τις κατηγορίες σχιζοφρένεια, κατάθλιψη, νεύρωση, επιδημιολογία και δημόσια υγεία. Παρουσιάζονται ακόμη και σε πολύ ποσοτικούς τομείς όπως οι επιστήμες της μηχανικής και ο βιομηχανικός έλεγχος ποιότητας.

*Οι κατηγορικές μεταβλητές είναι πολλών τύπων [49]και χωρίζονται σε τέσσερα είδη:*

**Ονομαστικές μεταβλητές**(Nominal variable), όπου οι κατηγορίες δεν έχουν κάποια διάταξη, αλλά δηλώνουν μόνο το όνομα της κατηγορίας. Για παράδειγμα, το φύλο έχει ως δυνατές απαντήσεις τις κατηγορίες άντρας και γυναίκα, έτσι για να μετρηθεί το σύνολο των παρατηρήσεων στην κάθε κατηγορία θα μπορούσαν οι απαντήσεις να αντικατασταθούν από δύο αριθμούς αντίστοιχα, όπως 0 και 1, χωρίς αυτοί οι αριθμοί να υποδηλώνουν κάποιου είδους διάταξη.

**Διατεταγμένες μεταβλητές** (Ordinal variable), όπου οι κατηγορίες έχουν κάποια διάταξη, αλλά δεν υποδηλώνουν ίση διαφορά από κατηγορία σε κατηγορία. Για παράδειγμα, το μορφωτικό επίπεδο ενός ατόμου έχει ως δυνατές απαντήσεις βασικές υποχρεωτικές σπουδές, μεταλυκειακές σπουδές, πτυχίο ΑΕΙ/ΤΕΙ, μεταπτυχιακός τίτλος σπουδών διδακτορικός τίτλος σπουδών. Οι απαντήσεις θα μπορούσαν να αντικατασταθούν με αριθμούς από το 1 μέχρι το 5, όπου το ένα δηλώνει το κατώτερο επίπεδο και το 5 το ανώτερο μορφωτικό επίπεδο.

**Ισοδιαστημικές μεταβλητές** (Interval variable), όπου οι κατηγορίες δεν είναι μόνο διατεταγμένες αλλά παρουσιάζουν και την ίδια διαφορά μεταξύ τους. Για παράδειγμα, στις ηλικιακές ομάδες εάν οι πρώτες δύο κατηγορίες είναι ηλικιακές ομάδες [25-30) και [30-35), τότε η διαφορά των διαστημάτων είναι ίδια ανεξάρτητα της ηλικίας.

**Αναλογικές μεταβλητές** (Ratio variable), όπου έχουν όλες τις ιδιότητες της ισοδιαστημικής μεταβλητής και επιπλέον ορίζεται ένα σημείο αναφοράς το μηδέν, το οποίο δηλώνει ότι το χαρακτηριστικό που μετριέται για την συγκεκριμένη μεταβλητή, στο σημείο αυτό δεν υπάρχει. Τέτοια παραδείγματα αναλογικής μεταβλητής είναι η ταχύτητα, το βάρος, η επιτάχυνση.

## 7.3 Περιγραφική Στατιστική Κατηγορικών Δεδομένων

Τα κατηγορικά δεδομένα μπορούν να αναπαρασταθούν από διακριτές τιμές (0,1,2,3,..), οι οποίες δηλώνουν το όνομα της κατηγορίας και δεν έχουν αριθμητική σημασία.΄Έτσι, η μόνη δυνατότητα επεξεργασίας κατηγορικών δεδομένων είναι η δημιουργία **Πινάκων Συχνοτήτων** και **Σχετικών Συχνοτήτων**, καθώς και η οπτικοποίηση αυτών των μεγεθών από διάφορες γραφικές παραστάσεις.

Για να περιγράψουμε την κατανομή των δεδομένων χρησιμοποιούμε μέτρα θέσης και μέτρα διασποράς, αλλά στα ποιοτικά δεδομένα δεν είναι αποδεκτά αυτά τα περιγραφικά μέτρα που χρησιμοποιούν αριθμητικούς υπολογισμούς. Είναι αναμενόμενο πως μεγέθη, όπως ο αριθμητικός μέσος, δεν δίνει αξιόπιστα αποτελέσματα. Βέβαια για ποιοτικές μεταβλητές που είναι αριθμημένες σε διατεταγμένη κλίμακα (Ordinal scale), η διάμεσος είναι ένα κατάλληλο μέτρο θέσης, αφού για τον υπολογισμό της χρειάζεται τα δεδομένα να είναι σε διάταξη. Επίσης για τέτοιου είδους μεταβλητές είναι αποδεκτοί στατιστικοί έλεγχοι που βασίζονται σε διατεταγμένα δεδομένα, όπως για παράδειγμα, προσημικός έλεγχος (sign test).

### 7.3.1 Γραφικές Αναπαραστάσεις

Στη στατιστική οι γραφικές αναπαραστάσεις είναι πολύ χρήσιμες για την ανάλυση των δεδομένων. Με τη βοήθεια στατιστικών πακέτων (R,SPSS,SAS) υπάρχουν διάφορα γραφήματα τα οποία είτε απεικονίζουν την κατανομή των μεταβλητών, είτε μας παρέχουν πληροφορίες για τη διάμεσο, είτε για τυχόν ακραίες τιμές.

Τέτοια γραφήματα είναι τα παρακάτω:

- **Ραβδόγραμμα συχνοτήτων και σχετικών συχνοτήτων(Barchart)**

Σε ένα τέτοιο διάγραμμα τα δεδομένα του πίνακα συχνοτήτων αναπαρίστανται από ράβδους. Ο άξονας των τεταγμένων αποτελείται είτε από τις συχνότητες είτε από τις σχετικές συχνότητες της μεταβλητής και ο άξονας των τετμημένων από τις τιμές της. Σε αυτή την περίπτωση, το ραβδόγραμμα έχει σχεδιασθεί με οριζόντιο προσανατολισμό, αντίστοιχα μπορεί α σχεδιασθεί και με κατακόρυφο προσανατολισμό.

- **Σημειόγραμμα - (Dot diagram)**

Στο σημειόγραμμα απεικονίζονται όλα τα δεδομένα ως κουκίδες. Χρησιμοποιείται σε ποιοτικά́ δεδομένα αλλά́ έχει καλύτερο αποτέλεσμα για μικρό́ δείγμα.

- **Κυκλικό́ διάγραμμα - (Pie Chart)**

Στο κυκλικό́ διάγραμμα απεικονίζονται οι συχνότητες ή οι σχετικές συχνότητες των τιμών της μεταβλητής σε μορφή́ κυκλικών τομέων.

- **Διάγραμμα συχνοτήτων (ή σχετικών συχνοτήτων) - (Line diagram)**

Το διάγραμμα συχνοτήτων είναι αντίστοιχο με το ραβδόγραμμα, με τη διαφορά ότι αντί για ορθογώνια έχουμε κάθετα ευθύγραμμα τμήματα στις αντίστοιχες τιμές της μεταβλητής.

- **Διάγραμμα Pareto - (Pareto Chart )**

Το διάγραμμα Pareto είναι ένα ραβδόγραμμα το οποίο τοποθετεί́ τις κατηγορίες μιας κατηγορικής μεταβλητής σε φθίνουσα σειρά́ ανάλογα με τη συχνότητα που συγκεντρώνει η κάθε κατηγορία, όπου υπάρχει και δεξιός κάθετος άξονας ο οποίος δείχνει το ποσοστό́ που καταλαμβάνουν οι κατηγορίες αθροιστικά́.

Το Ιστόγραμμα Συχνοτήτων (Frequency Histogram) και Σχετικών Συχνοτήτων καθώς και το Πολύγωνο Συχνοτήτων (Frequency Polygon) και Σχετικών Συχνοτήτων είναι

διαγράμματα που εφαρμόζουν σε ποσοτικά δεδομένα, ενώ τα παραπάνω διαγράμματα που αναφέρθηκαν είναι καταλληλά και για ποιοτικά δεδομένα. Δύο ακόμα σημαντικά διαγράμματα που υποδεικνύουν πληροφορίες για τις τιμές του δείγματος είναι το θηκόγραμμα (Boxplot) και το φυλλογράφημα (steam-leaf plot). Το θηκόγραμμα δείχνει τη διάμεσο καθώς και τις ακραίες τιμές του δείγματος και το φυλλογράφημα δείχνει όλες τις τιμές του δείγματος και πόσες φορές εμφανίστηκαν στο δείγμα. Προφανώς, τα δυο τελευταία διαγράμματα είναι χρήσιμα σε ποσοτικά δεδομένα.

### 7.3.2 Έλεγχος ανεξαρτησίας σε κατηγορικές μεταβλητές

Για να ελεγχθεί η υπόθεση ότι δυο κατηγορικές μεταβλητές, οι οποίες αφορούν τον ίδιο πληθυσμό, είναι μεταξύ τους ανεξάρτητες χρησιμοποιείται ο $\chi^2$ έλεγχος και για να εφαρμοσθεί ο συγκεκριμένος έλεγχος χρησιμοποιούνται πίνακες συνάφειας.

### 7.3.3 Πίνακες Συνάφειας

Οι **πίνακες συνάφειας** είναι πίνακες οι οποίοι ταξινομούν σταυρωτά τα χαρακτηριστικά δυο κατηγορικών μεταβλητών. Ένας τέτοιος πίνακας μπορεί να αποτελείται από πολλές γραμμές και στήλες αναλόγως με το πόσες κατηγορίες έχουν οι δυο μεταβλητές. Συχνά στην βιβλιογραφία είναι γνωστοί και ως πίνακες διπλής εισόδου.

## 7.4 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση **(Logistic regression)** αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Y με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή Y συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δυο τιμές) στοχεύετε η πρόβλεψη της έκβασης αυτής από ένα πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι ονομαστικές, τακτικές ή ποσοτικές.

Ορισμένες εφαρμογές χρησιμοποιούν λογική παλινδρόμηση για να μοντελοποιήσουν την πιθανότητα ότι ένα υποκείμενο είναι αξιόπιστο.

Η επιλογή μοντέλου για λογιστική παλινδρόμηση αντιμετωπίζει τα ίδια προβλήματα όπως για τη συνηθισμένη παλινδρόμηση. Η διαδικασία επιλογής καθίσταται δυσκολότερη καθώς ο αριθμός των επεξηγηματικών μεταβλητών αυξάνεται λόγω της ταχείας αύξησης των πιθανών επιδράσεων και αλληλεπιδράσεων.

Υπάρχουν δύο ανταγωνιστικοί στόχοι:

- Το μοντέλο πρέπει να είναι αρκετά σύνθετο ώστε να ταιριάζει καλά στα δεδομένα.
- Από την άλλη πλευρά, θα πρέπει να είναι απλή η ερμηνεία, η εξομάλυνση και όχι η υπερφόρτωση των δεδομένων.

Οι περισσότερες μελέτες έχουν σχεδιαστεί για να απαντούν σε ορισμένες ερωτήσεις. Αυτές οι ερωτήσεις καθοδηγούν την επιλογή των όρων του μοντέλου. Οι επιβεβαιωτικές αναλύσεις χρησιμοποιούν στη συνέχεια ένα περιορισμένο σύνολο μοντέλων. Για παράδειγμα, μια υπόθεση μελέτης σχετικά με ένα αποτέλεσμα μπορεί να δοκιμαστεί συγκρίνοντας μοντέλα με και χωρίς αυτό το αποτέλεσμα. Για μελέτες που είναι διερευνητικές και όχι επιβεβαιωτικές, η αναζήτηση μεταξύ πιθανών μοντέλων μπορεί να παρέχει ενδείξεις σχετικά με τη δομή εξάρτησης και να εγείρει ερωτήματα για μελλοντική έρευνα.

Υπάρχουν πολλές διαδικασίες επιλογής μοντέλων, καμία από τις οποίες δεν είναι πάντα καλύτερη. Μία μεταβλητή μπορεί να φαίνεται ότι έχει μικρή επίδραση επειδή επικαλύπτεται σημαντικά με άλλους προγνωστικούς παράγοντες στο μοντέλο, η οποία και η ίδια προβλέπεται καλά από τους άλλους προγνωστικούς παράγοντες. Η διαγραφή ενός τέτοιου πλεονάζοντος προγνωστικού μπορεί να είναι χρήσιμη, για παράδειγμα, για τη μείωση τυποποιημένων σφαλμάτων άλλων εκτιμώμενων αποτελεσμάτων[49].

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική), στη δε δεύτερη αποκλειστικά ποσοτική. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και b γίνεται με τη μέθοδο των ελαχίστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδό του λόγου πιθανοφάνειας (μέθοδος συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα), δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των

παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν.

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι:

1. **Δίτιμη** ή **δυαδική ή διχοτομική** (binary) ή **διμερής** εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν.
2. **Τακτική** (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ, στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος.
3. **Ονομαστική (Nominal)** ή **πολυώνυμή** (polynomial) ή **πολυχοτομική** (polychotomus) **ή κατηγορική αδιαβάθμητη** (non-ordered categorical) ή **πολυμερής** μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. ο χαρακτηρισμός ενός τρόφιμου ως τραγανού, μαλακού, εύθρυπτου ή του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ.

Η λογιστική παλινδρόμηση επινοήθηκε ως εναλλακτική επιλογή της γραμμικής διακριτικής ανάλυσης για την ταξινόμηση των στοιχείων (ονομαστικών ή τακτικών) της εξαρτημένης, με ευρεία απήχηση σε πολλά διαφορετικά επιστημονικά πεδία και κυρίως στην ιατρική και τις κοινωνικές επιστήμες.

Χαρακτηριστικά, χρησιμοποιείται στην πρόβλεψη της:

- εμφάνισης ή μη μιας νόσου (π.χ. διαβήτης) από ένα σύνολο διαφορετικών χαρακτηριστικών του πάσχοντος ατόμου (ηλικία, φύλο, αιματολογικά, ηλεκτροκαρδιογράφημα κτλ.)

- επιλογής ενός πολιτικού κόμματος με βάση την καταγραφή των δημογραφικών στοιχειών των πολιτών, όπως είναι η ηλικία, φύλο, φυλή, τόπος διαμονής, εισόδημα, προηγούμενη ψηφοφορία.
- πιθανότητας αποτυχίας μιας διεργασίας παραγωγής προϊόντος σε ένα εργοστάσιο τροφίμων.
- πρόβλεψη της πρόθεσης αγοράς ενός αγαθού από έναν καταναλωτή (έρευνα αγοράς).
- πιθανότητας αθέτησης από δανειολήπτη της αποπληρωμής του δανείου του.

Επειδή, στην παρούσα διπλωματική εργασία η εξαρτημένη μεταβλητή ερμηνεύεται από περισσότερους από έναν παράγοντα, το μοντέλο που δημιουργείται είναι το πολλαπλό λογιστικό μοντέλο.

*Πολλαπλό λογιστικό μοντέλο:*

Είναι γεγονός ότι στις επιστήμες της υγείας δεν υπάρχει εξάρτηση μόνο από έναν παράγοντα, αλλά από πολλούς. Είναι λοιπόν αναγκαία η εκτίμηση υποδειγμάτων που να προβλέπουν την πιθανότητα, ένα άτομο να εμφανίσει τη διερευνώμενη νόσο (στην παρούσα διπλωματική εργασία είναι ο καρκίνος στις κύστες των ωοθηκών) μέσω πολλών μεταβλητών.

Ορισμός της Εξαρτημένης και των Ανεξάρτητων μεταβλητών

Κάθε μεταβλητή συνήθως επηρεάζεται από περισσότερες από μία μεταβλητές. Συνεπώς τα περισσότερα μοντέλα που μελετώνται από τους ερευνητές περιλαμβάνουν τη ανάλυση παλινδρόμησης για μια εξαρτημένη μεταβλητή όχι προς μία ανεξάρτητη μεταβλητή αλλά περισσότερες. Συνήθως τα προβλήματα που έχουν να αντιμετωπίσουν οι ερευνητές είναι πολυπαραγοντικά και όχι μονοπαραγοντικά. Αυτό σημαίνει ότι αρκετοί είναι οι παράγοντες που θα πρέπει να συμπεριληφθούν στο μοντέλο ώστε να ερμηνεύσουν τη σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων στον καλύτερο βαθμό. Ως εξαρτημένη μεταβλητή στο μοντέλο πολλαπλής παλινδρόμησης ορίζεται αυτή της οποίας τις μεταβολές θέλουμε να ερμηνεύσουμε αξιολογώντας τις τιμές των υπόλοιπων ανεξάρτητων μεταβλητών. Κατά τον ορισμό των ανεξάρτητων μεταβλητών πρέπει να είμαστε πολύ προσεκτικοί. Δεν πρέπει ο

αριθμός τους να είναι πολύ μεγάλος ειδικά αν το μέγεθος του δείγματος που έχουμε είναι μικρός αριθμός.

### 7.5 Βιοστατιστική

Η **βιοστατιστική** είναι η εφαρμογή των στατιστικών μεθόδων στη βιολογία και, συνηθέστερα, στην ιατρική. Επειδή οι ερευνητικές υποθέσεις και τα σενάρια στις επιστήμες της βιολογίας και της ιατρικής είναι ποικίλες, η βιοστατιστική συμπεριλαμβάνει γενικότερα κάθε ποσοτική, και όχι μόνο στατιστική, προσέγγιση που μπορεί να χρησιμοποιηθεί για να απαντήσει σε ερευνητικά ερωτήματα ή να ελέγξει την ορθότητα επιστημονικών θεωριών, υποθέσεων και σεναρίων. Ο Σχεδιασμός και η ανάλυση των κλινικών δοκιμών είναι ίσως η περισσότερο γνωστή εφαρμογή των στατιστικών μεθόδων στη ιατρική.

Πολλοί επιστήμονες της στατιστικής δυσανασχετούν για το διαχωρισμό της βιοστατιστικής από τη Στατιστική με την έννοια ότι ένας στατιστικός μπορεί να χειριστεί και Ιατρικά προβλήματα. Όμως η πράξη στις χώρες της Ευρώπης και τις Η.Π.Α έδειξε ότι χρειάζονται εξειδικευμένες γνώσεις (τόσο στην ορολογία όσο και στην ανάλυση) που ένας στατιστικός δεν γνωρίζει. Αυτό, σε συνδυασμό με αυξημένη ζήτηση Βιοστατιστικών επιστημόνων από Ιατρικές και φαρμακευτικές εταιρείες, οδήγησε στη δημιουργία εντατικών μεταπτυχιακών προγραμμάτων βιοστατιστικής ή ιατρικής στατιστικής σε πολλά πανεπιστήμια της Ευρώπης και των ΗΠΑ εκ των οποίων αρκετά είναι άμεσα συνδεδεμένα με τη διαδικασία της παραγωγής και της Ιατρικής έρευνας.

## 8. Η μηχανική μάθηση και η στατιστική μάθηση

Σύμφωνα με τον **Larry Wasserman** και τα δύο ασχολούνται με το ίδιο ερώτημα: πώς μπορούμε να μάθουμε από τα δεδομένα.

Ο **Robert Tibshirani**, ένας στατιστικός και ειδικός μηχανικής μάθησης στο Πανεπιστήμιο του Στάνφορντ, καλεί την μηχανική μάθηση "δοξασμένη στατιστική".

Σήμερα, τόσο η μηχανική μάθηση όσο οι στατιστικές τεχνικές χρησιμοποιούνται στην αναγνώριση προτύπων, στην ανακάλυψη της γνώσης και στην εξόρυξη δεδομένων. Και οι δύο αυτές μέθοδοι επικεντρώνονται στο να "τραβήξουν" γνώσεις ή ιδέες μέσα από τα δεδομένα. Όμως, οι μέθοδοί τους επηρεάζονται από τις εγγενείς διαφορές στις κουλτούρες από τις οποίες προέρχονται.

***Σχετίζονται μεταξύ τους, σίγουρα. Αλλά οι γονείς τους είναι διαφορετικοί.***

*Η μηχανική μάθηση είναι ένα πεδίο της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης. Ασχολείται με δομικά συστήματα που μπορούν να μάθουν από τα δεδομένα, αντί να ασχολείται με ρητά προγραμματισμένες οδηγίες. Ένα στατιστικό μοντέλο, από την άλλη πλευρά είναι ένα πεδίο των μαθηματικών. Η μηχανική μάθηση είναι συγκριτικά ένα νέο πεδίο.*

Η φτηνή υπολογιστική ισχύς και η διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων επέτρεψε στους επιστήμονες να εκπαιδεύσουν τους υπολογιστές για να μάθουν από την ανάλυση των δεδομένων. Αλλά, η στατιστική μοντελοποίηση υπήρχε πολύ πριν την εφεύρεση των υπολογιστών.

***Μεθοδολογικές διαφορές μεταξύ μηχανικής μάθηση και στατιστικής***

Η διαφορά μεταξύ των δύο είναι ότι η μηχανική μάθηση εστιάζει στη βελτιστοποίηση και την απόδοση σε σχέση με το συμπέρασμα που είναι αυτό που ανησυχεί την στατιστική.

Το παρακάτω παράδειγμα δείχνει το πώς ένας στατιστικός και ένας ειδικός στην μηχανική μάθηση θα περιγράψουν το αποτέλεσμα του ίδιου μοντέλου:

- **Ειδικός μηχανικής μάθησης**: "Το μοντέλο είναι 85% ακριβές στην πρόβλεψη του Υ δεδομένου των α, β και γ ".
- **Στατιστικός**: "το μοντέλο είναι 85% ακριβές στην πρόβλεψη του Υ δεδομένου των α, β και γ και είμαι 90% σίγουρος ότι αν ξανακάνεις το πείραμα θα επιτευχθεί το ίδιο αποτέλεσμα ".

Η μηχανική μάθηση δεν απαιτεί προηγούμενες παραδοχές σχετικά με τις υποκείμενες σχέσεις μεταξύ των μεταβλητών. Απλά πρέπει να "ρίξουμε" όλα τα δεδομένα που έχουμε στον υπολογιστή, και ο αλγόριθμος επεξεργάζεται τα δεδομένα και ανακαλύπτει τα πρότυπα, με τα οποία μπορούμε να κάνουμε προβλέψεις για το νέο σύνολο των δεδομένων. Η μηχανική μάθηση αντιμετωπίζει έναν αλγόριθμο σαν ένα μαύρο κουτί (black box), για όσο διάστημα αυτό λειτουργεί. Γενικά εφαρμόζεται σε υψηλά σύνολα διαστάσεων των δεδομένων (δηλαδή πολλές μεταβλητές και παρατηρήσεις), όσο περισσότερα είναι τα δεδομένα που έχετε, τόσο πιο ακριβής θα είναι η πρόβλεψή σας.

Σε αντίθεση, ο ρόλος του στατιστικού είναι να καταλάβει πώς συλλέχθηκαν τα δεδομένα, τις στατιστικές ιδιότητες των εκτιμητών (p-value, αμερόληπτες εκτιμήτριες), την υποκείμενη κατανομή του πληθυσμού που μελετά και τα είδη των ιδιοτήτων που θα περίμενε κανείς αν έκανε το πείραμα πολλές φορές . Θα πρέπει να ξέρει ακριβώς τι κάνει για να καταλήξει σε παραμέτρους που θα παρέχουν την ικανότητα πρόβλεψης. Οι τεχνικές στατιστικής μοντελοποίησης εφαρμόζονται συνήθως σε σύνολα χαμηλών διαστάσεων των δεδομένων, δηλαδή λιγότερες μεταβλητές και παρατηρήσεις.

Μπορεί να φαίνεται ότι η μηχανική μάθηση και η στατιστική μοντελοποίηση είναι δύο διαφορετικοί κλάδοι της προγνωστικής μοντελοποίησης. Η διαφορά μεταξύ των δύο έχει μειωθεί σημαντικά την τελευταία δεκαετία. Και οι δύο κλάδοι έχουν μάθει ο ένας από τον άλλο πολύ και θα συνεχίσουν να έρχονται πιο κοντά στο μέλλον.

Όμως, η κατανόηση της συνεργασίας και η γνώση των διαφορών τους επιτρέπει στους εκπαιδευόμενους της μηχανικής μάθησης και στους στατιστικούς να επεκτείνουν τις γνώσεις τους και να εφαρμόζουν ακόμη και τις μεθόδους εκτός του τομέα της ειδικότητας τους. Αυτή είναι η έννοια της «επιστήμης των δεδομένων» , η οποία έχει ως στόχο να γεφυρώσει το

χάσμα. Η συνεργασία και η επικοινωνία μεταξύ αυτών των δύο συναρπαστικών κλάδων που βασίζονται στα δεδομένα, μας επιτρέπει να λαμβάνουμε καλύτερες αποφάσεις που τελικά θα επηρεάσουν θετικά τον τρόπο που ζούμε.

# 9. Εργαλείο ανάλυσης - R studio

## 9.1 Εισαγωγή

Το R Studio είναι ένα ελεύθερο και ανοιχτού κώδικα ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) για το R, μια γλώσσα προγραμματισμού για στατιστική και γραφικά. Το R Studio ιδρύθηκε από τον JJ Allaire, δημιουργό της γλώσσας προγραμματισμού ColdFusion. Ο Hadley Wickham είναι ο επικεφαλής επιστήμονας στο RStudio.

Διατίθεται σε δύο εκδόσεις: R studio Desktop, όπου το πρόγραμμα εκτελείται τοπικά ως κανονική εφαρμογή επιφάνειας εργασίας και RStudio Server, το οποίο επιτρέπει την πρόσβαση στο R Studio χρησιμοποιώντας ένα πρόγραμμα περιήγησης ιστού ενώ εκτελείται σε έναν απομακρυσμένο διακομιστή Linux. Οι προ-συσκευασμένες διανομές της επιφάνειας εργασίας R Studio είναι διαθέσιμες για Windows, macOS και Linux.

Το R Studio διατίθεται σε ανοιχτού κώδικα και εμπορικές εκδόσεις και εκτελείται στην επιφάνεια εργασίας (Windows, macOS και Linux) ή σε πρόγραμμα περιήγησης συνδεδεμένο με το R Studio Server ή το RStudio Server Pro (Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE και SLES).

Είναι εύκολα επεκτάσιμο μέσω λειτουργιών, ενώ η κοινότητα R σημειώνεται για τις ενεργές συνεισφορές της σε πακέτα. Το R λόγω της κληρονομιάς S, έχει ισχυρότερες αντικειμενοστραφείς εγκαταστάσεις προγραμματισμού από τις περισσότερες στατιστικές γλώσσες υπολογιστών. Η επέκταση του R διευκολύνεται επίσης από τους κανόνες λεξικού του πεδίου εφαρμογής. Πολλές από τις τυπικές λειτουργίες του R είναι γραμμένες στο ίδιο το R, γεγονός που καθιστά εύκολο για τους χρήστες να ακολουθήσουν τις αλγοριθμικές επιλογές που έχουν γίνει. Γράφεται στη γλώσσα προγραμματισμού C, C ++, Python, Java, .NET και χρησιμοποιεί το πλαίσιο Qt για το γραφικό περιβάλλον χρήστη του.

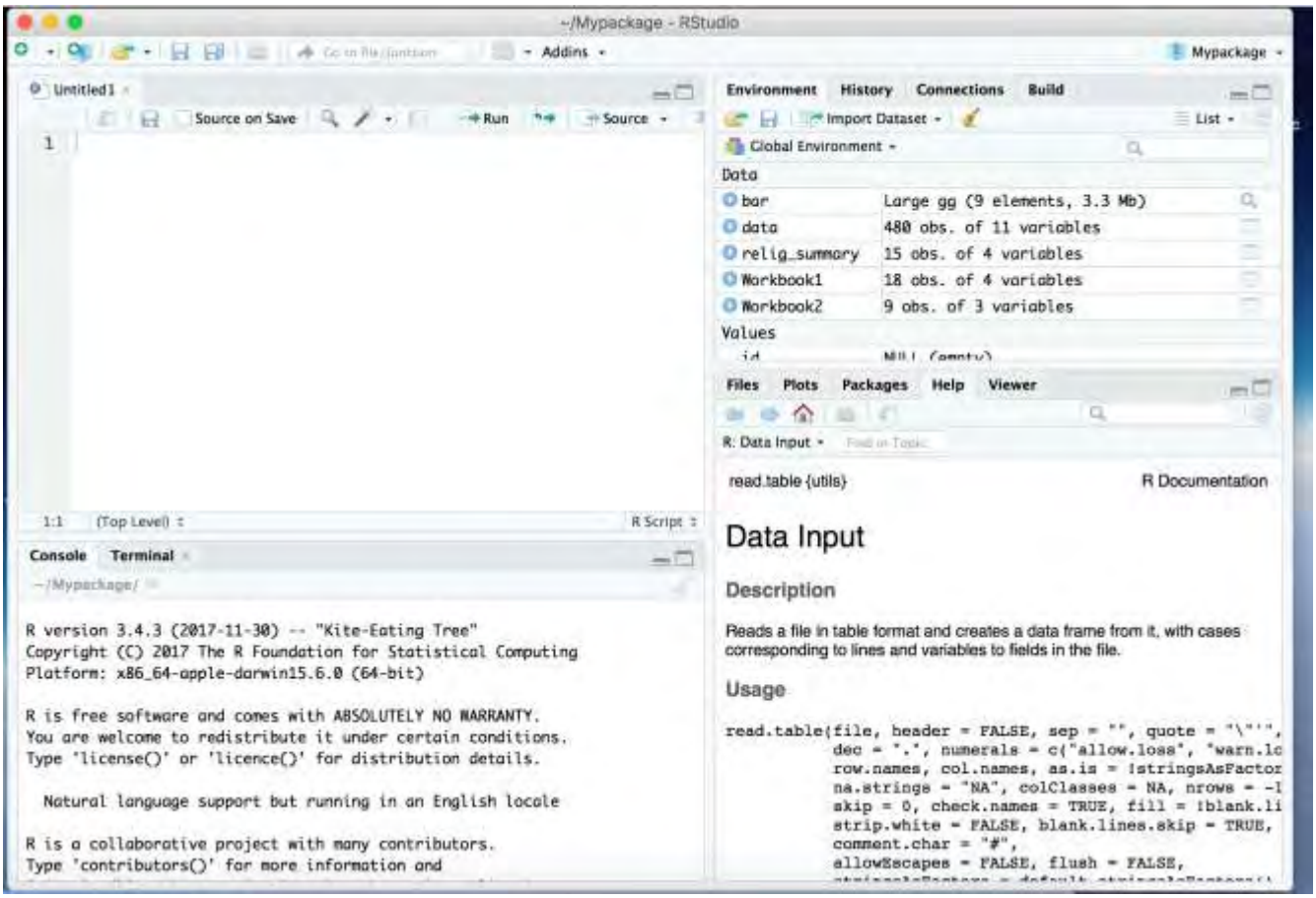
Μια άλλη δύναμη του R είναι τα στατικά γραφικά, τα οποία μπορούν να παράγουν γραφήματα ποιότητας δημοσίευσης, συμπεριλαμβανομένων των μαθηματικών συμβόλων. Δυναμικά και διαδραστικά γραφικά διατίθενται μέσω πρόσθετων πακέτων.

Το R έχει το Rd, τη δική του μορφή τεκμηρίωσης, όπως το LaTex, το οποίο χρησιμοποιείται για την παροχή περιεκτικής τεκμηρίωσης, τόσο on-line σε διάφορες μορφές όσο και σε έντυπη μορφή.

Η R είναι μία γλώσσα προγραμματισμού και ένα περιβάλλον ελεύθερου λογισμικού για στατιστικούς υπολογιστές και γραφικά που υποστηρίζεται από το Ίδρυμα Στατιστικής Πληροφορικής. Η γλώσσα R χρησιμοποιείται ευρέως μεταξύ των στατιστικών και των δεδομένων για την ανάπτυξη στατιστικού λογισμικού και την ανάλυση δεδομένων.

Η R και οι βιβλιοθήκες της εφαρμόζουν μια ευρεία ποικιλία στατιστικών και γραφικών τεχνικών, συμπεριλαμβανομένης της γραμμικής και μη γραμμικής μοντελοποίησης, κλασσικών και στατιστικών δοκιμών, ανάλυσης χρονοσειρών, ταξινόμησης, ομαδοποίησης και άλλων[50].

### 9.2 Το γραφικό περιβάλλον της R




# 10. Συλλογή και Ανάλυση δεδομένων με χρήση R studio

## 10.1 Στατιστική Ανάλυση

Το χρονικό διάστημα 2005-2017, συλλέχθηκαν δεδομένα από το Στρατιωτικό Νοσοκομείο Αθηνών, για 480 γυναίκες με κύστη ωοθήκης, από τις οποίες 340 ήταν στην αναπαραγωγική ηλικία και 140 στην εμμηνόπαυση. (**Πίνακας 1**)

| **Δημογραφικά Δεδομένα** | **Υπο-ομάδες** | **%Καλοήθεια** | **% Κακοήθεια** |
|---|---|---|---|
| **Ορμονικό Προφίλ** | Αναπαραγωγική Ηλικία | 76.4 (321 /420) | 31.7 (19/ 60) |
| | Εμμηνόπαυση | 23.5 (99 / 420) | 68.3 (41 / 60) |
| **Τόκος** | Χωρίς Παιδιά | 47.8 (201 / 420) | 21.6 (13 / 60) |
| | Παιδιά | 52.2 (219 / 420) | 78.3 (47 / 60) |

| | | | |
|---|---|---|---|
| **Προσωπικό ιστορικό Υστερεκτομής** | Αρνητικό | 96.4 (405 / 420) | 100 (60 / 60) |
| | Θετικό | 3.5 (15 / 420) | 0 (0 /60) |
| **Προσωπικό ιστορικό Ωοθηκεκτομής** | Αρνητικό | 87.8 (405 /461) | 12.5 (58 / 461) |
| | Θετικό | 73.7 (14 /19) | 15.8 (3 / 19) |
| **Ορμονική Θεραπεία** | Αρνητικό | 85.9 (361 / 420) | 95 (57 / 60) |
| | Θετικό | 14 (59 / 420) | 5 (3 / 60) |
| **Οικογενειακό ιστορικό Καρκίνου** | Αρνητικό | 87.8 (346 / 394) | 12.2 (48 / 394) |
| | Θετικό | 86.1 (74 / 86) | 14 (12 / 86) |
| **Οικογενειακό ιστορικό Καρκίνου των Ωοθηκών και/ή του Μαστού** | Αρνητικό | 87.5 (387 / 442) | 12.4 (55 / 442) |
| | Θετικό | 86.8 (33 / 38) | 13.2 (5 / 38) |
| **Προσωπκό ιστορικό Καρκίνου του Μαστού** | Αρνητικό | 88.2 (417 / 473) | 11.8 (56 / 473) |
| | Θετικό | 42.9 (3 / 7) | 57.1 (4 / 7) |
| **Προσωπικό ιστορικό Καρκίνου των Ωοθηκών** | Αρνητικό | 87.5 (420 / 480) | 12.5 (60 / 480) |
| | Θετικό | 0 | 0 |

**Πίνακας 1**

**Πίνακας 1:** Κατανομή των δημογραφικών δεδομένων των γυναικών που έχουν προσληφθεί σε σχέση με ιστολογικά αποδεδειγμένη καλοήθεια ή κακοήθεια των ωοθηκών.

Λαμβάνοντας υπόψη την ιστολογική διάγνωση, ο καρκίνος των ωοθηκών διαγιγνώσκεται στο 68,3% των γυναικών που βρίσκονται στην εμμηνόπαυση και στο 31,7% των γυναικών που είναι στην αναπαραγωγική ηλικία. Το 78,3% των γυναικών με ένα τουλάχιστον παιδί έχει παρατηρηθεί ότι έχουν προσβληθεί από καρκίνο των ωοθηκών σε σύγκριση με το 21,6% των ασθενών με καρκίνο του γεννητικού καρκίνου. Δεν έχουν εντοπιστεί κακοήθεις βλάβες των ωοθηκών μεταξύ των γυναικών που υποβλήθηκαν σε υστερεκτομή στο παρελθόν. Στην ομάδα των γυναικών που ανέφεραν προσωπικό ιστορικό ωοθηκεκτομής, οι κακοήθεις περιπτώσεις των ωοθηκών παρατηρήθηκαν στο 15,8% αυτών. Το 14% των γυναικών που διαγνώστηκαν με καλοήθη μάζα των ωοθηκών έλαβαν ορμονική θεραπεία τους τελευταίους έξι μήνες πριν την υπερηχογραφική εξέταση. Μόνο το 5% περίπου των ασθενών που έχουν προσβληθεί από καρκίνο των ωοθηκών, χρησιμοποίησαν ορμονική θεραπεία την ίδια χρονική περίοδο. Το οικογενειακό ιστορικό τουλάχιστον 1 κρούσματος καρκίνου ή πιο συγκεκριμένα του καρκίνου των ωοθηκών και / ή του μαστού δεν φαίνεται να επηρεάζει σημαντικά τη διάγνωση της κακοήθειας των ωοθηκών στο δείγμα μας (12,4% και 13,2% αντίστοιχα). Το προσωπικό ιστορικό της καρκινικής βλάβης του μαστού σχετίζεται σε μεγάλο βαθμό με την εμφάνιση του καρκίνου των ωοθηκών (57,1%).

**Πίνακας 2:** Επιπολασμός καλοηθών και κακοηθών ωοθηκών σε σχέση με τις ενδείξεις υπερηχογραφίας.

| **Ενδείξεις** | **% καλοήθεια** | **% κακοήθεια** |
|---|---|---|
| **Σηνήθη ετήσια εξέταση** | 36.8 (154 / 420) | 5 (3 / 60) |
| **Μετά από προηγούμενη ταυτοποίηση της κύστης των ωοθηκών** | 19.6 (82 / 420) | 6.6 (4 / 60) |

| | | |
|---|---|---|
| **Αίσθημα διάχυτου κοιλιακού πόνου** | 16.8 (72 / 420) | 18.3 (11 / 60) |
| **Επείγουσα περίπτωση έντονου κοιλιακού πόνου** | 8.1 (34 / 420) | 3.3 (2 / 60) |
| **Εμμηνορυσιακές ανωμαλίες** | 9.8 (41 / 420) | 25 (15 / 60) |
| **Υποβοηθούμενη διαγνωστική επεξεργασία (Υπογονιμότητα)** | 1.9 (8 / 420) | 1.7 (1 / 60) |
| **Μετεωρισμός** | 2.4 (10 / 420) | 30 (18 / 60) |
| **Συμπτώματα του ουροποιητικού συστήματος** | 1.7 (7 / 420) | 3.3 (2 / 60) |
| **Άλλα** | 2.9 (12 / 420) | 6.6 (4 / 60) |

**Πίνακας 2**

Ο μετεωρισμός, οι ανωμαλίες της εμμήνου ρύσεως και ο βαθύς κοιλιακός πόνος αποτελούν τα κύρια συμπτώματα που ενίσχυαν αρχικά τις γυναίκες που εξετάστηκαν (30%, 25% και 18,3% αντίστοιχα, πίνακας 2). Σε μικρότερο βαθμό, επείγοντα περιστατικά έντονου κοιλιακού πόνου (3,3%), συμπτωμάτων ούρων (3,3%) και διαγνωστικής διαδικασίας υπογονιμότητας (1,7%) οδήγησαν επίσης τους ασθενείς σε υπερηχογραφική εξέταση και συνεπώς ανακάλυψη κακοήθων ωοθηκικών βλαβών. Σχεδόν το 7% των ασθενών με καρκίνο (άλλα) διαμαρτυρήθηκαν αρχικά για δυσμηνόρροια, εύκολη κόπωση και βαθύ πόνο στην πλάτη. Αξίζει να σημειωθεί ότι γυναίκες χωρίς προκαταρκτικά συμπτώματα ή γυναίκες που παρακολουθήθηκαν για μάζα ωοθηκών έχουν διαγνωσθεί ότι έχουν προσβληθεί από καρκίνο των ωοθηκών (5% και 6,6% αντίστοιχα).

# 11. Αποτελέσματα και συμπεράσματα

**<u>Παράγοντες που χρησιμοποιήθηκαν και εφαρμογή αλγορίθμων</u>**

### 11.1 Οι παράγοντες που χρησιμοποιήθηκαν σε κώδικα

<u>Import data (Εισαγωγή Δεδομένων)</u>

```
> data<-read.csv(file.choose(),header=T)
```

> data

**Υπάρχουν 480 σειρές και 11 στήλες:**

dim(data)

[19] 480  11

**Οι επικεφαλίδες των μεταβλητών είναι:**

names(data)

 [19] "Hormonal.profile"

 [19] "Parity"

 [3] "History.of.hysterectomy"

 [19] "History.of.ovariectomy"

 [5] "Hormonal.therapy"

 [6] "Family.history.of.Ca.cases"

 [7] "Family.history.of.ovarian.and.or.breast.cancer"

 [19] "personal.history.of.breast.cancer"

 [19] "personal.history.of.ovarian.cancer"

[10] "Benign.Malignant"

[11] "Symptoms"

**Η παρακάτω εντολή μας δείχνει ότι πρόκειται για κατηγορικά δεδομένα:**

> data$Benign.Malignant=as.factor(data$Benign.Malignant)

> is.factor(data$Benign.Malignant)

[19] TRUE

> class(data$Benign.Malignant)

[19] "factor"

> class(data$Parity)

[19] "factor"

**Μας δείχνει τι περιέχει η κάθε μεταβλητή:**

levels(data$Benign.Malignant)

[19] "0" "1"

Το 0 αντιστοιχεί στο Benign(καλοήθεια)και το 1 στο Malignant(κακοήθεια)

**Μας δείχνει συνοπτικά τις μεταβλητές και το τι περιέχουν:**

>Summary(data)

| Hormonal.profile | Parity | History.of.hysterectomy | History.of.ovariectomy | Hormonal.therapy | Family.history.of.Ca. cases |
|---|---|---|---|---|---|
| postmenopausal:140 | No:214 | negative:465 | negative:463 | negative:418 | negative:394 |
| premenopausal :340 | yes:266 | positive: 15 | positive: 17 | positive: 62 | positive: 86 |

| Family.history.of.ovarian.and.or.breast.cancer | Personal.history.of.breast.cancer | Personal.history.of.ovarian.cancer | Benign.Malignant |
|---|---|---|---|
| negative:442 | negative:473 | Negative:480 | 0:420 |
| positive: 38 | positive: 7 | | 1:60 |

| Symptoms |
|---|
| routine yearly examination :156 |
| following previously identified ovarian cyst : 86 |
| feeling of diffuse deep abdominal pain : 82 |
| menstrual irregularities : 56 |
| urgent case of intense abdominal pain : 36 |
| bloating : 29 |
| Other : 35 |

## 11.2 Λογιστική Παλινδρόμηση-Logistic regession

>model.data<-
glm(data$Benign.Malignant~Hormonal.profile+Parity+History.of.hysterectomy+History.of.ovariectomy+Hormonal.therapy+Family.history.of.Ca.cases+Family.history.of.ovarian.and.or.breast.cancer+personal.history.of.breast.cancer+Symptoms,family=binomial(link='logit'),data=data)
> summary(model.data)

*Εμφανίζει:*

Call:

glm(formula = data$Benign.Malignant ~ Hormonal.profile + Parity +
History.of.hysterectomy + History.of.ovariectomy + Hormonal.therapy +
Family.history.of.Ca.cases + Family.history.of.ovarian.and.or.breast.cancer +
personal.history.of.breast.cancer + Symptoms, family = binomial(link = "logit"),
data = data)
(Οι παραπάνω σειρές μας λένε ποιο σύνολο δεδομένων χειριζόμαστε, τις ετικέτες της απόκρισης και τις επεξηγηματικές μεταβλητές και το είδος του μοντέλου που προσαρμόζουμε( που στην περίπτωση μας είναι το δυαδικό logit) και τον τύπο του αλγόριθμου βαθμολόγησης για την εκτίμηση παραμέτρων)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.9092 | -0.4329 | -0.2065 | -0.1206 | 3.0815 |

Coefficients:

| | Estimate |
|---|---|
| (Intercept) | -16.0769 |
| Hormonal.profilepremenopausal | -1.7956 |
| Parityyes | 1.0822 |
| History.of.hysterectomypositive | -16.1351 |
| History.of.ovariectomypositive | -0.3361 |
| Hormonal.therapypositive | -0.7758 |
| Family.history.of.Ca.casespositive | 0.3444 |
| Family.history.of.ovarian.and.or.breast.cancerpositive | -0.3050 |
| personal.history.of.breast.cancerpositive | 1.2233 |
| Symptomsbloating | 16.6409 |
| Symptomsfeeling of diffuse deep abdominal pain | 14.4700 |
| Symptomsfollowing previously identified ovarian cyst | 13.1333 |
| Symptomsmenstrual irregularities | 15.6274 |
| Symptomsothers | 15.2594 |
| Symptomsroutine yearly examination | 12.9530 |
| Symptomssubfertility diagnostic workup | 14.2425 |
| Symptomsurgent case of intense abdominal pain | 13.0860 |
| Symptomsurinary system's symptoms | 14.5526 |
| | Std. Error |

| | |
|---|---|
| (Intercept) | 3956.1804 |
| Hormonal.profilepremenopausal | 0.3707 |
| Parityyes | 0.4192 |
| History.of.hysterectomypositive | 867.8923 |
| History.of.ovariectomypositive | 1.1485 |
| Hormonal.therapypositive | 0.6075 |
| Family.history.of.Ca.casespositive | 0.4579 |
| Family.history.of.ovarian.and.or.breast.cancerpositive | 0.6462 |
| personal.history.of.breast.cancerpositive | 0.9706 |
| Symptomsbloating | 3956.1804 |
| Symptomsfeeling of diffuse deep abdominal pain | 3956.1804 |

| | |
|---|---|
| Symptomsfollowing previously identified ovarian cyst | 3956.1804 |
| Symptomsmenstrual irregularities | 3956.1804 |
| Symptomsothers | 3956.1804 |
| Symptomsroutine yearly examination | 3956.1804 |
| Symptomssubfertility diagnostic workup | 3956.1805 |
| Symptomsurgent case of intense abdominal pain | 3956.1805 |
| Symptomsurinary system's symptoms | 3956.1805 |
| | z value |
| (Intercept) | -0.004 |
| Hormonal.profilepremenopausal | -4.844 |
| Parityyes | 2.581 |
| History.of.hysterectomypositive | -0.019 |
| History.of.ovariectomypositive | -0.293 |
| Hormonal.therapypositive | -1.277 |
| Family.history.of.Ca.casespositive | 0.752 |
| Family.history.of.ovarian.and.or.breast.cancerpositive | -0.472 |
| personal.history.of.breast.cancerpositive | 1.260 |
| Symptomsbloating | 0.004 |
| Symptomsfeeling of diffuse deep abdominal pain | 0.004 |
| Symptomsfollowing previously identified ovarian cyst | 0.003 |
| Symptomsmenstrual irregularities | 0.004 |
| Symptomsothers | 0.004 |
| Symptomsroutine yearly examination | 0.003 |
| Symptomssubfertility diagnostic workup | 0.004 |
| Symptomsurgent case of intense abdominal pain | 0.003 |
| Symptomsurinary system's symptoms | 0.004 |
| | Pr(>\|z\|) |
| (Intercept) | 0.99676 |
| Hormonal.profilepremenopausal | 1.27e-06 *** |
| Parityyes | 0.00984 ** |
| History.of.hysterectomypositive | 0.98517 |
| History.of.ovariectomypositive | 0.76982 |
| Hormonal.therapypositive | 0.20157 |
| Family.history.of.Ca.casespositive | 0.45198 |
| Family.history.of.ovarian.and.or.breast.cancerpositive | 0.63693 |
| personal.history.of.breast.cancerpositive | 0.20755 |
| Symptomsbloating | 0.99664 |
| Symptomsfeeling of diffuse deep abdominal pain | 0.99708 |
| Symptomsfollowing previously identified ovarian cyst | 0.99735 |
| Symptomsmenstrual irregularities | 0.99685 |
| Symptomsothers | 0.99692 |
| Symptomsroutine yearly examination | 0.99739 |
| Symptomssubfertility diagnostic workup | 0.99713 |

| | |
|---|---|
| Symptomsurgent case of intense abdominal pain | 0.99736 |
| Symptomsurinary system's symptoms | 0.99707 |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 361.70 on 479 degrees of freedom
Residual deviance: 229.61 on 462 degrees of freedom
AIC: 265.61

Number of Fisher Scoring iterations: 16

Αυτή είναι μια λίστα με τις πιθανότητες καταγραφής σε κάθε επανάληψη. (Η λογιστική παλινδρόμηση χρησιμοποιεί μέγιστη πιθανότητα, η οποία είναι μια επαναληπτική διαδικασία.) Η πρώτη επανάληψη (που ονομάζεται επανάληψη 0) είναι η πιθανότητα καταγραφής του μοντέλου "null", δηλαδή ένα μοντέλο χωρίς προγνωστικούς δείκτες. Στην επόμενη επανάληψη, ο πρόλογος συμπεριλαμβάνεται στο μοντέλο. Σε κάθε επανάληψη, η πιθανότητα καταγραφής αυξάνεται επειδή ο στόχος είναι να μεγιστοποιηθεί η πιθανότητα καταγραφής. Όταν η διαφορά μεταξύ των διαδοχικών επαναλήψεων είναι πολύ μικρή, το μοντέλο λέγεται ότι "συγκλίνει", το iterating σταματά και τα αποτελέσματα εμφανίζονται.

Η λογιστική παλινδρόμηση είναι κατάλληλη όταν η μεταβλητή απόκρισης είναι κατηγορική με δύο πιθανά αποτελέσματα (δηλ. Δυαδικά αποτελέσματα). Οι δυαδικές μεταβλητές μπορούν να αναπαρασταθούν χρησιμοποιώντας μία μεταβλητή δείκτη Yi, λαμβάνοντας τις τιμές 0 ή 1 και μοντελοποιημένες χρησιμοποιώντας μια δυαδική κατανομή με πιθανότητα P (Yi = 1) = i. Η λογιστική παλινδρόμηση μοντελοποιεί αυτή την πιθανότητα ως μια συνάρτηση μιας ή περισσοτέρων επεξηγηματικών μεταβλητών.

**Απόκλιση(Deviance)**

Η απόκλιση είναι ένα μέτρο της καλής προσαρμογής ενός γενικευμένου γραμμικού μοντέλου. Ή μάλλον, είναι ένα μέτρο κακής εφαρμογής - υψηλότεροι αριθμοί δείχνουν χειρότερη εφαρμογή. Το R αναφέρει δύο μορφές αποκλίσεων - τη μηδενική αποκλίση και την υπολειπόμενη απόκλιση. Η μηδενική απόκλιση δείχνει πόσο καλά υπολογίζεται η μεταβλητή απόκρισης από ένα μοντέλο που περιλαμβάνει μόνο την απόσταση (μεγάλη μέση τιμή).

**Null αποκλίσεις (Null Deviance)**

Η περίληψη του μοντέλου λέει:
Null αποκλίσεις: 361,70 σε 479 βαθμούς ελευθερίας

Όταν το μοντέλο περιλαμβάνει μόνο τον όρο αλληλεπίδρασης, τότε η απόδοση του μοντέλου διέπεται από μηδενική απόκλιση.

**Υπολειπόμενη απόκλιση (Residual deviance)**

Η περίληψη του μοντέλου λέει:

Υπολειπόμενη απόκλιση: 229,61 στους 462 βαθμούς ελευθερίας

Η υπολειπόμενη απόκλιση είναι χαμηλότερη (229,61) από την μηδενική απόκλιση (361,70). Η χαμηλότερη τιμή της υπολειπόμενης απόκλισης επισημαίνει ότι το μοντέλο έχει βελτιωθεί.

**Βαθμός ελευθερίας (Degree of freedom):**

Η περίληψη στην έξοδο λέει:
Null αποκλίσεις: 361,70 σε 479 βαθμούς ελευθερίας

Οι βαθμοί ελευθερίας για μηδενική αποκλίση ισούνται με το N-1, όπου N είναι ο αριθμός των παρατηρήσεων στο δείγμα δεδομένων. Εδώ N = 480, συνεπώς N-1 = 480-1 = 479

Η περίληψη στην έξοδο λέει:
Υπολειπόμενη απόκλιση: 229,61 στους 462 βαθμούς ελευθερίας

Οι βαθμοί ελευθερίας για υπολειπόμενη απόκλιση ισοδυναμούν με N-k-1, όπου k είναι ο αριθμός μεταβλητών και N είναι ο αριθμός παρατηρήσεων σε δείγμα δεδομένων. Εδώ N = 480, k = 9, συνεπώς N-k-1 = 480-9-1 = 470

Οι βαθμοί ελευθερίας που σχετίζονται με την μηδενική και υπολειπόμενη απόκλιση διαφέρουν κατά 9 (479-470), καθώς το μοντέλο έχει 9 μεταβλητές (Hormonal.profile, Parity, History of hysterectomy, History of ovariectomy, Hormonal therapy, Family history of Ca cases, Family history of ovarian and or breast cancer, personal history of breast cancer, Symptoms), έχουν υπολογιστεί εννιά πρόσθετες παράμετροι και επομένως έχουν καταναλωθεί εννιά πρόσθετοι βαθμοί ελευθερίας.

**AIC:**

Η περίληψη στην έξοδο λέει:
AIC: 265,61

Η πλήρης μορφή του είναι το κριτήριο πληροφοριών Akaike (AIC) παρέχει μια μέθοδο για την αξιολόγηση της ποιότητας του μοντέλου μέσω της σύγκρισης των σχετικών μοντέλων. Είναι βασισμένο στην Απόκλιση. Αυτό είναι χρήσιμο όταν έχουμε περισσότερα από ένα μοντέλα για να συγκρίνουμε την καλοσύνη της τοποθέτησης των μοντέλων. Πρόκειται για μια εκτίμηση μέγιστων πιθανοτήτων που επιβάλλεται για να αποφευχθεί η υπερφόρτωση. Μετράει την ευκαμψία των μοντέλων. Είναι ανάλογο με το προσαρμοσμένο $R^2$ σε πολλαπλή γραμμική παλινδρόμηση όπου προσπαθεί να αποτρέψει το ενδεχόμενο να συμπεριληφθούν οι άσχετες μεταβλητές πρόβλεψης. Η χαμηλή AIC του μοντέλου είναι καλύτερη από το μοντέλο που έχει υψηλότερο AIC.

Η περίληψη στην έξοδο λέει:
Αριθμός επαναλήψεων του Fisher: 16

Οι κλειστές εξισώσεις φόρμας μπορούν να χρησιμοποιηθούν για επίλυση για παραμέτρους γραμμικού μοντέλου, αλλά δεν μπορούν να χρησιμοποιηθούν για λογιστική παλινδρόμηση.
Μια επαναληπτική προσέγγιση γνωστή ως αλγόριθμος Newton-Raphson χρησιμοποιείται για αυτό. Ο αλγόριθμος βαθμολόγησης του Fisher είναι ένα παράγωγο της μεθόδου του Newton για την επίλυση αριθμητικών προβλημάτων μέγιστης πιθανότητας.

Λέει πώς εκτιμήθηκε το μοντέλο. Ο αλγόριθμος κοιτάζει γύρω του για να διαπιστώσει αν η προσαρμογή θα βελτιωθεί χρησιμοποιώντας διαφορετικές εκτιμήσεις. Αν βελτιωθεί τότε κινείται προς αυτή την κατεύθυνση και στη συνέχεια προσαρμόζεται πάλι στο μοντέλο. Ο αλγόριθμος σταματά όταν δεν μπορεί να πραγματοποιηθεί σημαντική πρόσθετη βελτίωση. "Αριθμός επαναλήψεων βαθμολόγησης Fisher" λέει "πόσες επαναλήψεις τρέχει αυτός ο αλγόριθμος πριν σταματήσει" .Αυτό είναι 16.

Το αστεράκι(*) δίπλα στην τιμή p-value μας δείχνει ότι είναι στατιστικά σημαντικές μεταβλητές.
Παρατηρούμε ότι οι μεταβλητές Parity και Hormonal profile είναι σημαντικές, ένω οι υπόλοιπες μη σημάντικες διότι, p>0.05.

Ας ελέγξουμε τους βασικούς όρους που χρησιμοποιούνται στη λογιστική παλινδρόμηση και στη συνέχεια να προσπαθήσουμε να βρούμε την πιθανότητα να πάρει "χαμηλή = 1" (δηλαδή ευκολία επιτυχίας).

Odds ratio=probability of success(p)/ probability of failure=probability of (target variable=1)/probability of (target variable=0) = p/(1-p)

Logit score:

$$\text{logit}(p) = \log(p/(1-p)) = b0 + b1*x1 + \ldots + bk*xk$$

Τώρα ας ακολουθήσουμε τα παρακάτω σημεία για να βρούμε την τελική πιθανότητα (μεταβλητή στόχου = 1 ή χαμηλή = 1) από το αποτέλεσμα logit:

1. Συντελεστής αντιστάθμισης – Intercept Coefficient (b0) = -16,0769
2. Συντελεστής Hormonal profile (b1) = - 1,7956
   Συντελεστής Parity (b2) = -1,0822

Ερμηνεία:
Η αύξηση της βαθμολογίας logit ανά μονάδα αύξησης του Hormonal profile είναι – 1,7956

Η αύξηση της βαθμολογίας logit ανά μονάδα αυξάνει την μεταβλητή Parity και είναι - 1,0822

3. p-value για την μεταβλητή Parity = 0.00984
Σύμφωνα με το z-test, η τιμή p είναι 0.00984 η οποία είναι συγκριτικά χαμηλή, υποδηλώνει ότι είναι απίθανο ότι δεν υπάρχει «καμία σχέση» μεταξύ μεταβλητής Parity και στόχου δηλαδή χαμηλή μεταβλητή.

4. p-value για την μεταβλητή Hormonal profile = 1,27e-06

Σύμφωνα με το z-test, η p-value είναι 1,27e-06 η οποία είναι συγκριτικά υψηλή και συνεπάγεται ότι είναι απίθανο να υπάρχει "οποιαδήποτε σχέση" μεταξύ της ηλικίας και της μεταβλητής στόχου.

Std. Err. :

Αυτά είναι τα τυπικά σφάλματα που σχετίζονται με τους συντελεστές. Το τυπικό σφάλμα χρησιμοποιείται για να ελεγχθεί εάν η παράμετρος είναι σημαντικά διαφορετική από 0. Διαιρώντας την εκτίμηση των παραμέτρων με το τυπικό σφάλμα, λαμβάνετε μια τιμή z-value. Τα τυπικά σφάλματα μπορούν επίσης να χρησιμοποιηθούν για να σχηματίσουν ένα διάστημα εμπιστοσύνης για την παράμετρο, όπως φαίνεται στις δύο τελευταίες στήλες του παραπάνω πίνακα.

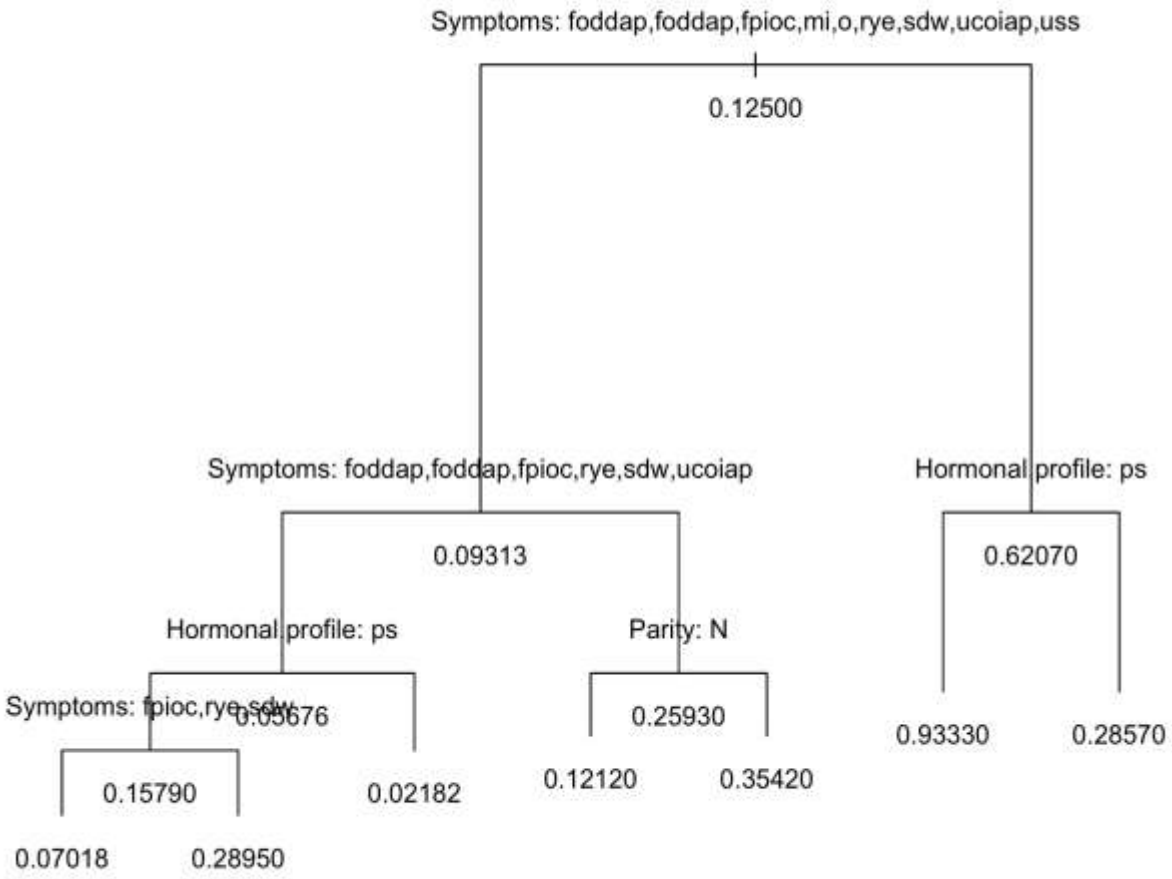
z and P>|z| :

Αυτές οι στήλες παρέχουν την τιμή z-value και την τιμή p-value που χρησιμοποιήθηκε για τη δοκιμή της μηδενικής υπόθεσης ότι ο συντελεστής (παράμετρος) είναι 0. Οι συντελεστές που έχουν τιμές p-value μικρότερες από 0,05 είναι στατιστικά σημαντικές.

### 11.3 Δέντρα Ταξινόμησης (Classification Tree)

```
>ecoli.tree1=tree(data$Benign.Malignant ~ Hormonal.profile + Parity +History.of.ovariectomy +
Hormonal.therapy + Family.history.of.Ca.cases + Family.history.of.ovarian.and.or.breast.cancer
+ personal.history.of.breast.cancer+Symptoms, data = data)

>plot(ecoli.tree1)
> text(ecoli.tree1, all = T, pretty = T)
```

Εμφανίζει:




Τα δέντρα απόφασης είναι μια εντελώς διαφορετική μέθοδος εκτίμησης των λειτουργικών μορφών σε σύγκριση με τη γραμμική παλινδρόμηση.

Για ένα δέντρο ταξινόμησης, προβλέπετε ότι κάθε παρατήρηση ανήκει στην πιο συχνή κατηγορία παρατηρήσεων εκπαίδευσης στην περιοχή στην οποία ανήκει.

Κατά την ερμηνεία των αποτελεσμάτων ενός δένδρου ταξινόμησης, συχνά ενδιαφερόμαστε όχι μόνο για την πρόβλεψη της τάξης που αντιστοιχεί σε μια συγκεκριμένη περιοχή τερματικού κόμβου αλλά και για τις τάξεις που ανήκουν στις παρατηρήσεις εκπαίδευσης που εμπίπτουν σε αυτή την περιοχή.

Η ρίζα ή ο κορυφαίος κόμβος του δέντρου είναι ο κόμβος απόφασης που χωρίζει το σύνολο δεδομένων χρησιμοποιώντας μια μεταβλητή ή ένα χαρακτηριστικό που έχει ως αποτέλεσμα την καλύτερη μετρική διαίρεση που αξιολογείται για κάθε υποσύνολο ή κλάση στο σύνολο δεδομένων που προκύπτει από διαίρεση. Το δέντρο αποφάσεων μαθαίνει με την

αναδρομική διάσπαση του συνόλου δεδομένων από τη ρίζα προς τα εμπρός (με άπληστους όρους κόμβου ανά κόμβο) σύμφωνα με τη μετρική διαίρεσης σε κάθε κόμβο απόφασης.

Η πιο σημαντική μεταβλητή είναι τα συμπτώματα- οι ενδείξεις (Symptoms) για τις οποίες η γυναίκα πηγαίνει στο γυναικολόγο.

Τα πιο σημαντικά είναι :

- Το αίσθημα διάχυτου κοιλιακού πόνου
- Μετά από προηγούμενη ταυτοποίηση της κύστης των ωοθηκών (Following previously identified ovarian cyst)
- Η συνήθη ετήσια εξέταση (Routine yearly examination)
- Η υποβοηθούμενη διαγνωστική επεξεργασία- Υπογονιμότητα (Subfertility diagnostic workup
- Η επείγουσα περίπτωση έντονου κοιλιακού πόνου.

Τα παραπάνω συμπτώματα σε συνδυασμό με το ότι η γυναίκα βρίσκεται στην εμμηνόπαυση (postmenopausal), δείχνουν πως πρόκειται για κακοήθη κύστη ωοθηκών.

### 11.4 Random Forest

>Boston.rf=randomForest(data$Benign.Malignant~Hormonal.profile+Parity+History.of.ovariectomy+Hormonal.therapy+Family.history.of.Ca.cases+Family.history.of.ovarian.and.or.breast.cancer + personal.history.of.breast.cancer+Symptoms, data = data)

>Boston.rf

>varImpPlot(Boston.rf)

Εμφανίζει:

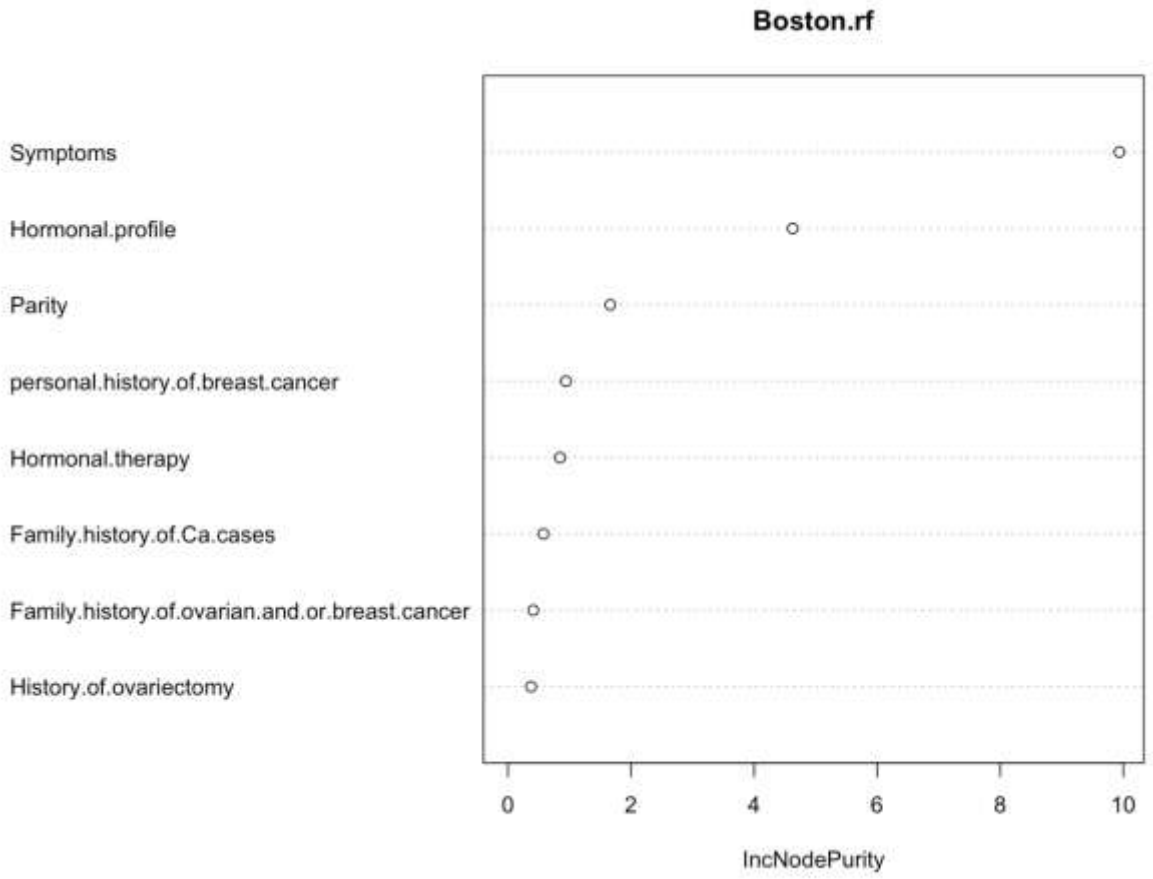
Call:

randomForest(formula=data$Benign.Malignant~Hormonal.profile+Parity+History.of.ovariectomy+Hormonal.therapy+Family.history.of.Ca.cases+
Family.history.of.ovarian.and.or.breast.cancer + personal.history.of.breast.cancer + Symptoms, data = data)

Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 2

Mean of squared residuals: 0.08388477
% Var explained: 23.31

Χρησιμοποιήθηκαν 500 δέντρα.
Το μοντέλο χρησιμοποίησε τυχαία 2 μεταβλητές για να καθορίσει τη διάσπαση.
Το ποσοστό σφάλματος OOB δημιουργείται χρησιμοποιώντας δέντρα που δεν ήταν κατάλληλα για την παρατήρηση. Εδώ έχουμε μια εκτίμηση OOB του ποσοστού σφάλματος 23%.




Αυτό μας λέει πόσο μειώνει η κάθε μεταβλητή τον μέσο δείκτη Gini, ένα μέτρο για το πόσο σημαντική είναι η μεταβλητή στο μοντέλο. Ουσιαστικά, εκτιμά την επίπτωση που έχει μια μεταβλητή στο μοντέλο, συγκρίνοντας τα ποσοστά ακρίβειας πρόβλεψης για τα μοντέλα με και χωρίς τη μεταβλητή. Μεγαλύτερες τιμές υποδεικνύουν μεγαλύτερη σημασία της μεταβλητής. Εδώ βλέπουμε ότι η μεταβλητή Symptoms είναι πιο σημαντική. Ακολουθούν οι μεταβλητές Hormonal.profile και Parity σε επίπεδο σημαντικότητας και τέλος οι υπόλοιπες, όπως φαίνεται στο διάγραμμα.

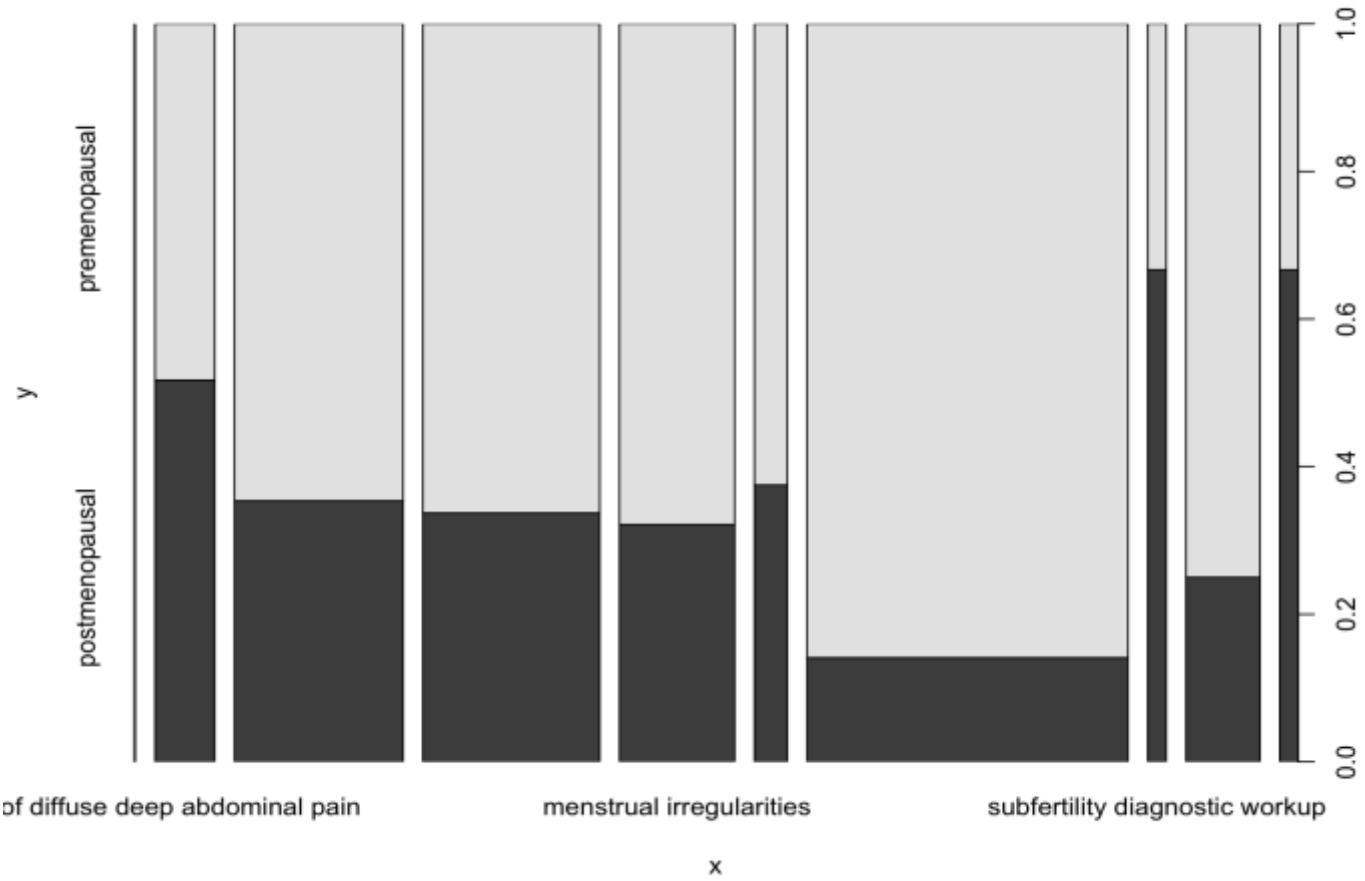
**11.5 SVM**

```
> library(e1071)
> model<-svm(Species~.,data=iris)
> x<-data$Benign.Malignant
> y<-data$Symptoms
> model<-svm(x,y)
> print(model)

Call:
svm.default(x = x, y = y)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  1

Number of Support Vectors:  408

> summary(model)

Call:
svm.default(x = x, y = y)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  1

Number of Support Vectors:  408

 ( 1 86 56 82 9 36 27 16 9 )


Number of Classes:  9

Levels:
  feeling of diffuse deep abdominal pain bloating feeling of diffuse deep abdominal pain following previously identified ovarian cyst menstrual irregularities others routine yearly examination subfertility diagnostic workup urgent case of intense abdominal pain urinary system's symptoms
```

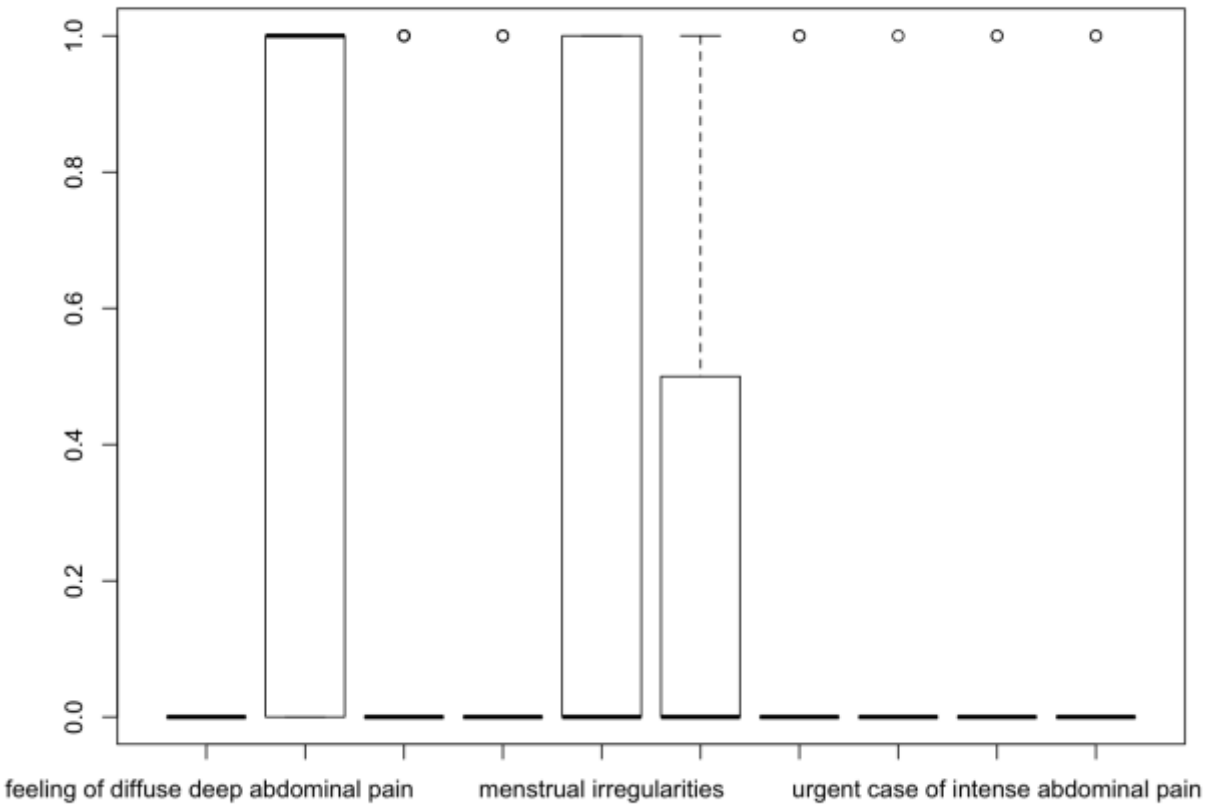
> pred<-predict(model,x)
> pred<-fitted(model)
> table(pred,x)

| | x | |
|---|---|---|
| pred | 0 | 1 |
| feeling of diffuse deep abdominal pain | 0 | 0 |
| bloating | 0 | 60 |
| following previously identified ovarian cyst | 0 | 0 |
| menstrual irregularities | 0 | 0 |
| others | 0 | 0 |
| routine yearly examination | 420 | 0 |
| subfertility diagnostic workup | 0 | 0 |
| urgent case of intense abdominal pain | 0 | 0 |
| urinary system's symptoms | 0 | 0 |

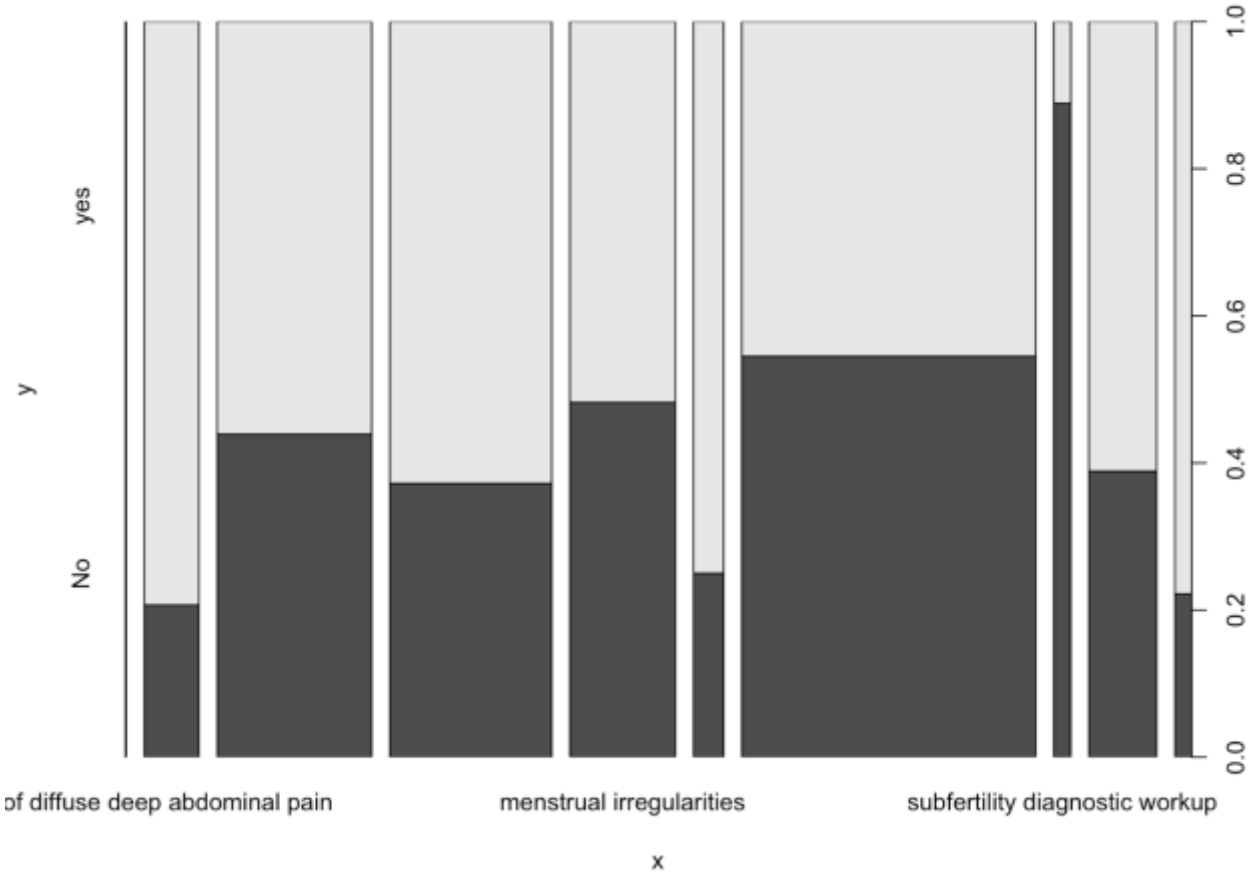
>pred<-predict(model,x,decision.values=TRUE)
> attr(pred,"decision.values")[1:5,]
> plot(data$Symptoms,data$Hormonal.profile)



>plot(data$Symptoms,data$Benign.Malignant)




> plot(data$Symptoms,data$Parity)




- svm () - Χρησιμοποιείται για την εκπαίδευση SVM.

Μπορεί να κάνει γενική παλινδρόμηση και ταξινόμηση, καθώς και εκτίμηση πυκνότητας.

Τα παρακάτω δεδομένα περιγράφουν μερικές παραμέτρους εισαγωγής της συνάρτησης svm ():

α) Data - Καθορίζει ένα προαιρετικό πλαίσιο δεδομένων που περιέχει τις μεταβλητές που υπάρχουν σε ένα μοντέλο. Όταν χρησιμοποιείται αυτή η παράμετρος, τότε δεν χρειάζεται να χρησιμοποιούνται οι παραμέτροι x και y. Παίρνουμε τις μεταβλητές από προεπιλογή από το περιβάλλον που καλείται 'svm'.

X – ένα διάνυσμα (αντικείμενο της κλάσης Matrix που παρέχεται από το πακέτο Matrix). Αντιπροσωπεύει τις περιπτώσεις του συνόλου δεδομένων και των αντίστοιχων ιδιοτήτων τους. Σε ένα διάνυσμα, οι σειρές αντιπροσωπεύουν τις περιπτώσεις και οι στήλες αντιπροσωπεύουν τις ιδιότητες.

b) Type(Τύπος) - Μπορούμε να χρησιμοποιήσουμε το SVM ως μηχάνημα ταξινόμησης, μηχάνημα παλινδρόμησης ή για ανίχνευση καινοτομίας. Ανάλογα με το αν το y είναι

παράγοντας ή όχι. Η προεπιλεγμένη ρύθμιση για τον τύπο είναι C-ταξινόμηση ή eps-παλινδρόμηση. Μπορεί να αντικατασταθεί με τη ρύθμιση μιας ρητής τιμής. Οι έγκυρες επιλογές είναι:

C-classification (ταξινόμηση)
Nu-classification (μη-ταξινόμηση)
One-classification (μία ταξινόμηση για ανίχνευση καινοτομίας)
eps-regression (παλινδρόμηση)
Nu-regression (μη-παλινδρόμηση)
Degree (βαθμός)

γ) Parameter (παράμετρος) - Απαιτείται για τον πυρήνα του τύπου πολυώνυμο

gamma - παράμετρος που απαιτείται για όλους τους πυρήνες εκτός από το γραμμικό (προεπιλογή: 1 / (διάσταση δεδομένων))
coef0 - απαραίτητη παράμετρος για τους πυρήνες τύπου πολυωνύμου και sigmoid (προεπιλογή: 0)
cost - το κόστος παραβίασης των περιορισμών (προεπιλογή: 1) -είναι η σταθερά «C» του όρου νομιμοποίησης στη διατύπωση Lagrange.

- predicate () - Χρησιμοποιώντας αυτή τη μέθοδο αποκτάται πρόβλεψη από το μοντέλο, καθώς και τιμές απόφασης από τους δυαδικούς ταξινομητές.

Η συνάρτηση πρόβλεψης - predicate () προβλέπει τιμές που βασίζονται σε μοντέλο που παράγεται από ένα SVM. Επιστρέφει τις ετικέτες της κλάσης σε περίπτωση ταξινόμησης με μια τιμή ιδιότητας μέλους ή τις τιμές απόφασης του ταξινομητή. Επιστρέφει επίσης ένα διάνυσμα προβλεπόμενων ετικετών για ένα πρόβλημα ταξινόμησης.

- plot () - Οπτικοποίηση δεδομένων, διανύσματα υποστήριξης και όρια απόφασης, εφόσον παρέχονται.

Η έξοδος από το svm model μας δείχνει ότι πρόκειται για C-classification.
Υπάρχουν 408 support vectors. (Υπάρχει μια διάκριση μεταξύ ενός μαλακού περιθωρίου και ενός σκληρού περιθωρίου svm. Αυτό ελέγχεται από την παράμετρο κόστους που επιβάλλει τον βαθμό στον οποίο επιτρέπεται στους φορείς υποστήριξης να παραβιάζουν τον περιορισμό περιθωρίου.)

Στην πρόβλεψη που κάνει το μοντέλο μας δείχνει ότι 420 γυναίκες πήγαν στο νοσοκομείο για την συνήθη ετήσια εξέταση και εμφάνισαν καλοήθη κύστη ωοθήκης, και 60 γυναίκες εισήχθησαν στο νοσοκομείο με φούσκωμα, στις οποίες βρέθηκε κακοήθεια.

Στα διαγράμματα έγινε σύγκριση της μεταβλητής Symptoms (που βρέθηκε σε προηγούμενα μοντέλα σημαντική) με τις μεταβλητές Hormonal.profile, Benign.Malignant και Parity. Τα αποτελέσματα δείχνουν ότι τα πιο σημαντικά συμπτώματα είναι το αίσθημα διάχυτου κοιλιακού πόνου, οι ανωμαλίες της εμμήνου ρύσεως, η υπογονιμότητα και η επείγουσα περίπτωση έντονου κοιλιακού πόνου, που σε συνδυασμό με το αν η γυναίκα βρίσκεται στην εμμηνόπαυση και έχει παιδιά αυξάνεται η πιθανότητα καρκίνου.

**11.6 Naïve Bayes**

```
> library(e1071)
> χ<-cbind(data$Benign.Malignant_train,data$Hormonal.profile_train)
> y<-cbind(data$Benign.Malignant_train,data$Symptoms_train)
> fit<-naiveBayes(data$Benign.Malignant~.,data=data)
> summary(fit)
```

| | Length | Class | Mode |
|---|---|---|---|
| apriori | 2 | table | numeric |
| tables | 10 | -none- | list |
| levels | 0 | -none- | NULL |
| call | 4 | -none- | call |

```
> fit<-naiveBayes(data$Symptoms~.,data=data)
> summary(fit)
```

| | Length | Class | Mode |
|---|---|---|---|
| apriori | 10 | table | numeric |
| tables | 10 | -none- | list |
| levels | 10 | -none- | character |
| call | 4 | -none- | call |

```
> fit<-naiveBayes(data$Parity~.,data=data)
> summary(fit)
```

| | Length | Class | Mode |
|---|---|---|---|
| apriori | 2 | table | Numeric |
| Tables | 10 | -none- | List |
| levels | 2 | -none- | character |
| call | 4 | -none- | call |

```
> print(fit)
```

Με αυτή την μεταβλητή, το μοντέλο δημιουργεί την υποθετική πιθανότητα για κάθε χαρακτηριστικό χωριστά (όπως φαίνεται παρακάτω). Έχουμε επίσης τις a priori πιθανότητες που υποδεικνύουν τη διανομή των δεδομένων μας.

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

| Y | |
|---|---|
| 0 | 1 |
| 0.875 | 0.125 |

Conditional probabilities:

| Hormonal.profile | | | | Parity | | |
|---|---|---|---|---|---|---|
| Y | postmenopausal | premenopausal | | Y | No | yes |
| 0 | 0.2357143 | 0.7642857 | | 0 | 0.4833333 | 0.5166667 |
| 1 | 0.6833333 | 0.3166667 | | 1 | 0.1833333 | 0.8166667 |

| History.of.hysterectomy | | | | History.of.ovariectomy | | |
|---|---|---|---|---|---|---|
| Y | negative | positive | | Y | negative | positive |
| 0 | 0.96428571 | 0.0357142 9 | | 0 | 0.96190476 | 0.0380952 4 |
| 1 | 1.0000000 0 | 0.0000000 0 | | 1 | 0.9833333 3 | 0.0166666 7 |

| Hormonal.therapy | | | | Family.history.of.Ca.cases | | |
|---|---|---|---|---|---|---|
| Y | negative | positive | | Y | negative | positive |
| 0 | 0.86190476 | 0.13809524 | | 0 | 0.8214286 | 0.1785714 |
| 1 | 0.93333333 | 0.06666667 | | 1 | 0.8166667 | 0.1833333 |

| Family.history.of.ovarian.and.or.breast.cancer | | |
|---|---|---|
| Y | negative | positive |
| 0 | 0.91904762 | 0.08095238 |
| 1 | 0.93333333 | 0.06666667 |

| Personal.history.of.breast.cancer | | | | Personal.history.of.ovarian.cancer | |
|---|---|---|---|---|---|
| Y | negative | positive | | Y | negative |
| 0 | 0.992857143 | 0.007142857 | | 0 | 1 |
| 1 | 0.933333333 | 0.066666667 | | 1 | 1 |

| Symptoms | | | | | | | |
|---|---|---|---|---|---|---|---|
| Y | feeling of diffuse deep abdominal pain | | bloating | | others routine yearly | | examination |
| 0 | 0.002380952 | | 0.026190476 | | 0.028571429 | | 0.364285714 |
| 1 | 0.000000000 | | 0.300000000 | | 0.066666667 | | 0.050000000 |

| Symptoms | | | |
|---|---|---|---|
| Y | feeling of diffuse deep abdominal pain | | following previously identified ovarian cyst |
| 0 | 0.169047619 | | 0.195238095 |
| 1 | 0.183333333 | | 0.066666667 |

| Symptoms | | | |
|---|---|---|---|
| Y | menstrual irregularities | | subfertility diagnostic workup |
| 0 | 0.097619048 | | 0.019047619 |
| 1 | 0.250000000 | | 0.016666667 |

| Symptoms | | |
|---|---|---|
| Y | urgent case of intense abdominal pain | urinary system's symptoms |
| 0 | 0.080952381 | 0.016666667 |

| | | |
|---|---|---|
| 1 | 0.033333333 | 0.033333333 |

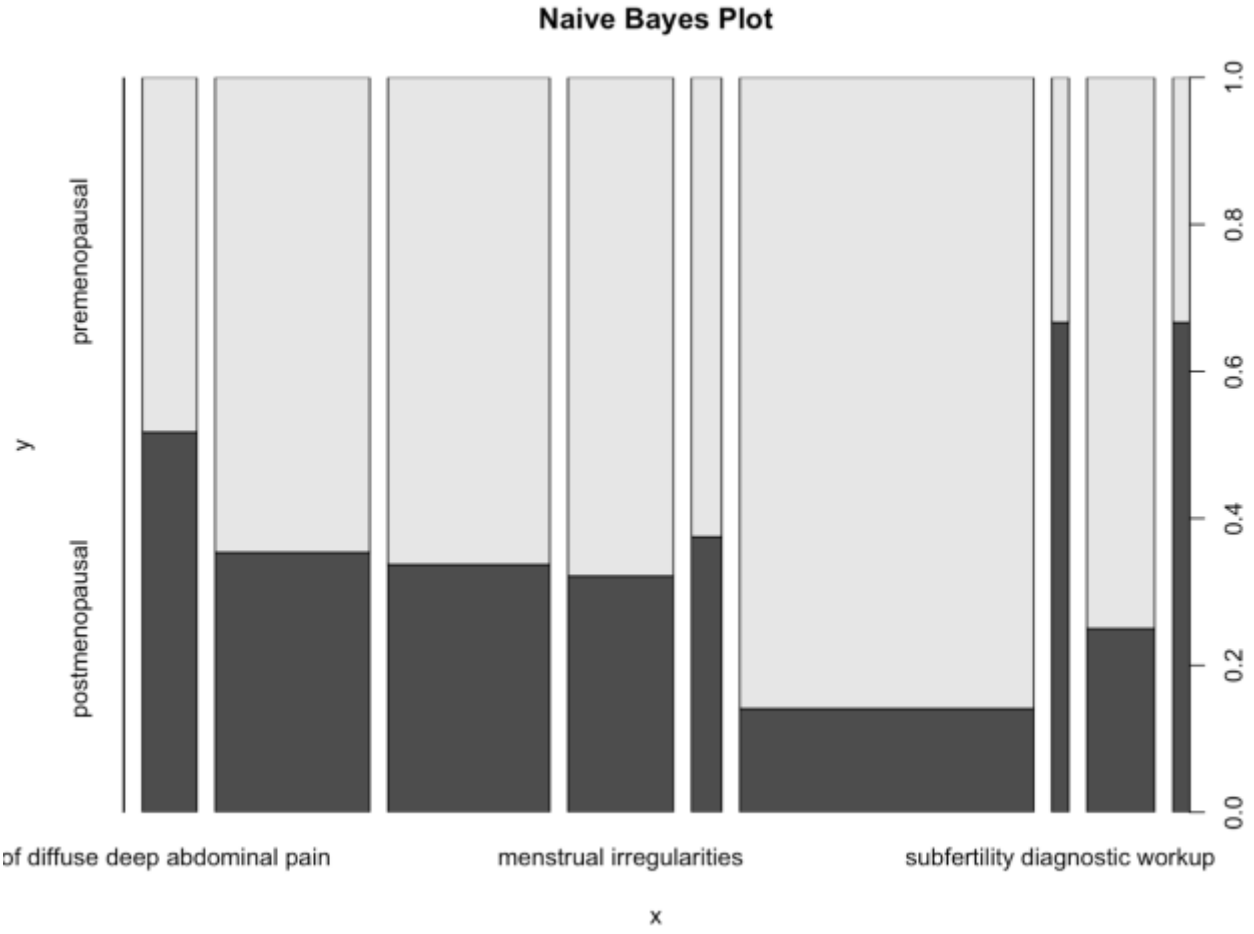
Ας υπολογίσουμε τον τρόπο με τον οποίο εκτελούμε τα δεδομένα

> table(predicted,data$Benign.Malignant)

| predicted | 0 | 1 |
|---|---|---|
| No | 336 | 0 |
| Yes | 84 | 60 |

Τα παραπάνω αποτελέσματα μας δείχνουν ότι μπορούμε να ταξινομήσουμε σωστά 336 από τις 420 περιπτώσεις καλοήθειας και 60 από τις 60 περιπτώσεις κακοήθειας.
Αυτό σημαίνει ότι η ικανότητα του Naive Bayes αλγορίθμου να προβλέψει 0 (καλοήθεια) είναι 80% και 1 (κακοήθεια) είναι 100%.

> plot(data$Symptoms,data$Hormonal.profile,main = "Naive Bayes Plot")




- Naïve Bayes, χρησιμοποιείται για να ταιριάζει με το μοντέλο Naive Bayes στο οποίο οι προβλεπόμενοι δείκτες θεωρούνται ανεξάρτητοι σε κάθε ετικέτα κλάσης.

Τα παρακάτω δεδομένα περιγράφουν μερικές παραμέτρους εισαγωγής της συνάρτησης naive bayes ():

a) X και Y, πλαίσιο δεδομένων με κατηγορικούς (χαρακτήρα / παράγοντα / λογικό) ή μετρικούς (αριθμητικούς) προγνωστικούς δείκτες.

   Cbind, μέθοδος που διαιρεί τις στήλες στο πλαίσιο δεδομένων και μετατρέπει τις στήλες χαρακτήρων σε παράγοντες.

   Data, πλαίσιο δεδομένων με κατηγορικά δεδομένα.

   Apriori είναι η προηγούμενη πιθανότητα για κάθε κλάση στο σετ εκπαίδευσης.

   Tables είναι η παράμετρος που αποθηκεύει τις πιθανότητες υπό όρους για κάθε συνδυασμό χαρακτηριστικών και κλάσεων.

   Levels είναι οι επιτρεπόμενες κλάσεις στο μοντέλο.
   Call είναι η κλήση που παρήγαγε το αντικείμενο που χρησιμοποιήσαμε.
b) Predict, η μέθοδος επιστρέφει έναν συντελεστή με τάξη που αντιστοιχεί στην μέγιστη πιθανότητα posterior ή ένα matrix με πιθανές posterior πιθανότητες για κάθε κατηγορία.
c) Plot, μέθοδος σχεδίασης για αντικείμενα κατηγορίας "naive_bayes" που έχουν σχεδιαστεί για μια γρήγορη ματιά στις περιθωριακές πιθανότητες των μεταβλητών πρόβλεψης δεδομένης της τάξης.

## 12.Τελικά Συμπεράσματα

Σκοπός αυτής της διπλωματικής ήταν να γίνει μια εισαγωγή στην ανάλυση δεδομένων με τεχνικές στατιστικής και μηχανικής μάθησης. Πρόκειται για κατηγορικά δεδομένα. Επιλέχθηκε το πρόγραμμα R studio, καθώς ενδείκνυται για εφαρμογές που περιέχουν statistical analysis.

Εκτός από την θεωρητική ανάλυση των μοντέλων που χρησιμοποιήθηκαν, ενδιαφέρον έχει η σύγκριση αυτών σε πείραμα με πραγματικά δεδομένα, καθώς και η αξιολόγηση των αποτελεσμάτων σε σύγκριση με θεωρητικές προσεγγίσεις για την εξαγωγή ενός τελικού πορίσματος. Υπάρχει αναλυτική περιγραφή όλων των βημάτων που ακολούθησα για την δημιουργία των αλγορίθμων.

Βασικό συμπέρασμα αποτελεί το γεγονός ότι ένα από τα σημαντικότερα βήματα στην όλη διαδικασία είναι η συλλογή και η προ επεξεργασία των δεδομένων. Δόθηκε μεγάλη προσοχή στο στάδιο καθαρισμού και φιλτραρίσματος των δεδομένων για να έχουμε τα τελικά δεδομένα που αποτελούν τα χαρακτηριστικά σύμφωνα με τα οποία έγινε η εκπαίδευση και η πρόβλεψη σε γυναίκες με κύστη ωοθήκης. Η ανάλυση των δεδομένων βοήθησε στη δημιουργία των αλγορίθμων με υψηλή ακρίβεια. Για την επιτυχία ενός καλού αποτελέσματος πρόβλεψης εφαρμόστηκαν τεχνικές μηχανικής μάθησης και πιο συγκεκριμένα, λογιστική παλινδρόμηση, δέντρα απόφασης, Random forest και οι αλγόριθμοι SVM και Naïve Bayes. Έγινε σύγκριση ως προς την αποτελεσματικότητα και την ταχύτητα των αλγορίθμων. Ο SVM αλγόριθμος δουλεύει καλύτερα με όσον το δυνατόν περισσότερα, αν είναι δυνατόν και όλα τα attributes του dataset και έχει ακριβείς προβλέψεις. Ο Naïve Bayes αλγόριθμος είναι ιδιαίτερα χρήσιμος για πολύ μεγάλα σύνολα δεδομένων. Επίσης είναι συγκρίσιμος σε απόδοση με το δέντρο αποφάσεων, έχει ελάχιστο ποσοστό σφάλματος σε σύγκριση με άλλους ταξινομητές και συγκριτικά με την λογιστική παλινδρόμηση είναι καλύτερος και χρησιμοποιεί λιγότερα δεδομένα εκπαίδευσης. Ο αλγόριθμος Random Forest έχει υψηλή ακρίβεια και καλή αξιοπιστία σε μεγάλες βάσεις δεδομένων. Επιβεβαιώθηκε μετά από πειραματισμούς και μετρήσεις πως οι τεχνικές που εφαρμόστηκαν μας έδωσαν τα ίδια αποτελέσματα. Πιο συγκεκριμένα, οι παράγοντες που βρέθηκαν να είναι στατιστικά σημαντικοί στην εμφάνιση κακοήθειας παρατηρήθηκαν στις γυναίκες που βρίσκονταν στην εμμηνόπαυση, είχαν τουλάχιστον ένα παιδί, είχαν προσωπικό ιστορικό καρκίνου του μαστού, δεν λάμβαναν ορμονική θεραπεία, η οποία προστατεύει από τον καρκίνο και σε συγκεκριμένα συμπτώματα που παρουσίασαν. Αυτά είναι ο μετεωρισμός, οι ανωμαλίες της εμμήνου ρύσεως, το αίσθημα διάχυτου και έντονου κοιλιακού πόνου. Επίσης κακοήθεια παρατηρήθηκε σε μεγάλο ποσοστό σε γυναίκες με υπογονιμότητα και γυναίκες που έκαναν τη συνήθη ετήσια εξέταση και είτε γνώριζαν από προηγούμενο έλεγχο την ύπαρξη κύστης ωοθηκών είτε όχι. Τα αποτελέσματα συνάδουν με την διεθνή βιβλιογραφία του χώρου της γυναικολογίας.

Αξίζει να σημειωθεί ότι το ποσοστό επιτυχίας εξαρτάται από την ποιότητα του δείγματος εκπαίδευσης αλλά και του δείγματος ελέγχου. Τα δεδομένα που ελέγχθηκαν ήταν τα δημογραφικά δεδομένα των γυναικών που έχουν προσληφθεί σε σχέση με ιστολογικά αποδεδειγμένη καλοήθεια ή κακοήθεια των ωοθηκών και οι ενδείξεις – συμπτώματα που παρουσίασαν οι γυναίκες με κύστη ωοθήκης. Από τα αποτελέσματα φαίνεται πως αποτέλεσαν

ιδανικές περιπτώσεις πινάκων για τους αλγορίθμους, όσον αφορά την απόδοση και την συμπεριφορά στο συγκεκριμένο σύνολο δεδομένων.

***Βιβλιογραφία***


1. Αλέξανδρος, Ά., Περιγραφική και εφαρμοσμένη ανατομική Β.Σπλάχνα. 1997: University studio press. 645.
2. Bannister, W.W.D., Gray's Anatomy 37th Edition. 1989, London: Churchill Livingstone. 1598.
3. Richard, S., Κλινική Ανατομική. 19912: Λίτσας Ιατρικές Εκδόσεις. 920.
4. Ellis, H., ClinicalAnatomy. 8 ed. 1992, Oxford: Blackwell Scientific Publications. 456.
5. Linder, H.H., Clinical Anatomy. 1989, San Francisco: Appleton & Lange. 650.
6. McMinn, R.M.H., Last's Anatomy. 1990, New York: Churchill Livingstone. 707.
7. Gougeon, A., Regulation of ovarian follicular development in primates: facts and hypotheses. Endocr Rev, 1996. 17(2): p. 121-55.
8. Yong, E.L., D.T. Baird, and S.G. Hillier, Mediation of gonadotrophin-stimulated growth and differentiation of human granulosa cells by adenosine-3',5'-monophosphate: one molecule, two messages. Clin Endocrinol (Oxf), 1992. 37(1): p. 51-8.
9. McNatty, K.P., et al., The production of progesterone, androgens, and estrogens by granulosa cells, thecal tissue, and stromal tissue from human ovaries in vitro. J Clin Endocrinol Metab, 1979. 49(5): p. 687-99.
10. Young, J.R. and R.B. Jaffe, Strength-duration characteristics of estrogen effects on gonadotropin response to gonadotropin-releasing hormone in women. II. Effects of varying concentrations of estradiol. J Clin Endocrinol Metab, 1976. 42(3): p. 432-42.
11. Yoshimura, Y., et al., The effects of proteolytic enzymes on in vitro ovulation in the rabbit. Am J Obstet Gynecol, 1987. 157(2): p. 468-75.
12. Lumsden, M.A., et al., Changes in the concentration of prostaglandins in preovulatory human follicles after administration of hCG. J Reprod Fertil, 1986. 77(1): p. 119-24.
13. Filicori, M., J.P. Butler, and W.F. Crowley, Jr., Neuroendocrine regulation of the corpus luteum in the human. Evidence for pulsatile progesterone secretion. J Clin Invest, 1984. 73(6): p. 1638-47.
14 le Nestour, E., et al., Role of estradiol in the rise in follicle-stimulating hormone levels during the luteal-follicular transition. J Clin Endocrinol Metab, 1993. 77(2): p. 439-42.
15 Kroon, E. and E. Andolf, Diagnosis and follow-up of simple ovarian cysts detected by ultrasound in postmenopausal women. Obstet Gynecol, 1995. 85(2): p. 211-4.
16 American College of, O. and Gynecologists, ACOG Practice Bulletin. Management of adnexal masses. Obstet Gynecol, 2007. 110(1): p. 201-14.
17. Aubuchon, M. and R.S. Legro, Polycystic ovary syndrome: current infertility management. Clin Obstet Gynecol, 2011. 54(4): p. 675-84.
18. Levine, D., et al., Simple adnexal cysts: the natural history in postmenopausal women. Radiology, 1992. 184(3): p. 653-9.

19. Te Linde, R.W.R.W.-., J.A. Rock, and H.W. Jones, III, 1942-, Te Linde's operative gynecology. 9th ed. / [edited by] John A. Rock, Howard W. Jones III. ed. 2003, United States: Philadelphia, Pa : Lippincott Williams & Wilkins, c2003.
20 .Malihe, H., T. Shamila, and M. Sara, Ovarian dermoid cyst. Prof Med 2010.
21 .Reid, B.M., J.B. Permuth, and T.A. Sellers, Epidemiology of ovarian cancer: a review. Cancer Biol Med, 2017. 14(1): p. 9-32.
22. Rakhlin, A. and K. Sridharan, Statistical Learning and Sequential Prediction. 2014: Autoedición. 261.
23. J.W, T., Exploratory Data Analysis. 1977: Addison-Wesley, Reading, MA. 712.
24 .Fayyad, U.M., et al., Advances in knowledge discovery and data mining. 1996: Menlo Park, CA: AAAI Press/ The MIT Press. 281.
25. Jackson, J.E., A user's guide to principal components. 1991: John Wiley, New York. 592.
26. Witten, I. and E. Frank, Data Mining:Practical Machine Learning Tools and Techniques, Second Edition: Morgan Kaufmann. 560.
27 Langley, P. and H.A. Simon, Applications of machine learning and rule induction. 1995, New York: ACM. 64.
28. Roxanne, C., G. Vernon, and L. Paul, Statistical modelling of key variables in social survey data analysis. 2016: p. 17.
29. Samuel, A.L., Computation & intelligence. 1959, USA: American Association for Artificial Intelligence Menlo Park. 414.
30. Mitchell, T.M., Machine Learning. 1997: McGraw-Hill Science/Engineering/Math. 432.
31. Han, J., M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Third Edition. 2012: Morgan Kaufmann. 744.
32. Vapnik, V.N., Statistical learning theory. 1998: : Wiley-Interscience. 768.
33. Zaki, M.J. and W. Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms. 2014: Cambridge University Press. 593.
34. Moore, A.W.; Available from: http://www.cs.cmu.edu/~awm/.
35. Dupa, R.O. and P.E. Hart, Pattern Classification and Scene Analysis. 1973, New York: Wiley. 512.
36. Cristianini, N. and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 2000: Cambridge University Press. 204.
37. Vapnik, V., The Nature of Statistical Learning Theory. 1995, New York: Springer-Verlag. 314.
38. Burges, C.J.C., A tutorial on support vector machines for pattern recognition. 1998.
39. Vapnik, V., S. Golowich, and A. Smo, Support vector method for function approximation, regression estimation, and signal processing. 1997: p. 287.
40. Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. 2009, New York: Springer-Verlag 745.
41. James, G., et al., An Introduction to Statistical Learning with Applications in R. 2013: Springer. 430.
42. Shalev-Shwartz, S. and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. 2014: Cambridge University Press. 410.
43. Elkan, C., Predictive analytics and Data mining. 2013: University of California.
44. J.P.Lewis, A Short SVM (Support Vector Machine) Tutorial. 2004.
45. Domingos, P. and M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss. 1997: p. 130.

46. Friedman, N., D. Geiger, and M. Goldszmidt, Bayesian network classifiers. 1997, USA: Kluwer Academic Publishers Hingham. 193.
47. Shalizi, C.R., Advanced Data Analysis from an Elementary Point of View. 2017: Cambridge University Press. 860.
48. Breiman, L., Random Forests. 2001: Kluwer Academic Publishers. 32.
49. Agresti, A., Categorical Data Analysis, second edition. 2002, New Jersey: John Wiley & Sons. 701.
50. John, V., Using R for introductory statistics. 2004: CRC Press. 432.